DATA-POP ALLIANCE
WHITE PAPER SERIES

Opportunities and Requirements for Leveraging Big Data for Official Statistics and the Sustainable Development Goals in Latin America

May 2016

DATA POP ALLIANCE

HARVARD HUMANITARIAN INITIATIVE

mit media lab

ODI

FLOWMINDER.ORG

Opportunities and Requirements for Leveraging Big Data for Official Statistics and the Sustainable Development Goals in Latin America

Julia Manske (Co-lead author and corresponding author)
David Sangokoya (Co-lead author), Data-Pop Alliance
Gabriel Pestre, Data-Pop Alliance
Emmanuel Letouzé, Data-Pop Alliance

May 2016

Contents

Foreword 1

Introduction 2

1 The State of Latin American NSOs: Overall Context and Concepts 7
- 1.1 The Role of National Statistical Offices in Latin America and the Caribbean . . . 7
- 1.2 The State of NSOs in LAC: Ongoing Challenges 9
- 1.3 Defining “Big Data” for Official Statistics and the SDGs 11

2 Engaging, Innovating, and Discovering Big Data in Latin America 17
- 2.1 Setting the Stage: the Burgeoning Ecosystem of Big Data 17
- 2.2 NSOs and Big Data: Trends in Latin America 21
- 2.3 Big Data for SDGs across the Wider Ecosystem of Actors 26
- 2.4 International Attempts to Use Big Data for Official Statistics and Development . 33

3 Challenges and Requirements for NSOs Engaging Big Data for SDGs 35
- 3.1 Institutional Barriers to Innovation and Change Management 35
- 3.2 Constraints to data access and completeness 37
- 3.3 Technical Challenges 39
- 3.4 Human Capacity Gaps 40
- 3.5 Methodological Challenges 42
- 3.6 Ethical and Political Risks 44

4 Towards a Regional Multi-Partner Roadmap for Leveraging Big Data for Official Statistics and the SDGs 47
- 4.1 Five Regional Trends Promoting Big Data Use in Latin America 47
- 4.2 Towards a Regional Multi-Partner Roadmap for Big Data: Building on Existing Regional Strengths and Opportunities 50

Glossary 56

Annexes 59

Figures

1 Map of National Statistical Offices in Latin America. 5
2 Selected Big Data Projects in Latin America. 6
3 Internet Use by Percentage of Population, 2006 and 2014 18
4 World Development of Digitalization 2013 21
5 Comparison of Self-Reported SDG Targets 34
6 Roadmap of Recommendations for Incorporating Big Data into NSO projects in LAC 52

Boxes

1 INEGI as Best Practice in the Region 12
2 Big Data vs. big data 15
3 Difference between Big Data and Open Data 16
4 NSOs in Latin America: Colombia's DANE Moderno 22
5 Twitter for Tourism Monitoring in Mexico 24
6 Maternal Morbidity and Remote Sensing of Malaria in Brazil 26
7 Using Google Trends to Nowcast Economic Data in Colombia 29

Tables

1 2013 Internet Usage and Population Statistics for Selected Countries in LAC . . . 19
2 Overview of Big Data Projects in Selected LAC NSOs 25
3 Overview of Big Data Ecosystem in LAC 30
4 Civil Society Organizations Working on Digital Rights 33

Annexes

1 Taxonomy and Examples of Big Data Sources 59
2 Uses of Big Data Towards Monitoring the SDGs 59
3 Analysis of Social Media Messages by Statistics Netherlands 64
4 Price Indices by Cavallo (MIT) 65
5 Luminosity Data as a Proxy for Economic Statistics 65
6 An Information System for Prices in Agriculture 66
7 Open Intelligence in Mexico 66
8 Leveraging Big Data Sources and Techniques Based on CDRs to Analyze Socioeconomic Outcomes and Processes in Colombia: The Cases of Public Safety and Social Development 67
9 Major Information Technology Components 68
10 Aggregated SWOT analysis for LAC NSOs and Big Data 69

Foreword

About this document

This document was produced as part of a World Bank-supported project implemented by Data-Pop Alliance in partnership with Colombia's *Departamento Administrativo Nacional de Estadística* (DANE). Data-Pop Alliance is a coalition on Big Data and development jointly created by the Harvard Humanitarian Initiative (HHI), the MIT Media Lab, and the Overseas Development Institute (ODI) to promote a people-centered Big Data revolution.

About the authors

This paper was written by the following authors:

- **Julia Manske** (Co-lead and corresponding author: jmanske@datapopalliance.org)
- **David Sangokoya** (Co-lead author), Data-Pop Alliance
- **Gabriel Pestre**, Data-Pop Alliance
- **Emmanuel Letouzé**, Data-Pop Alliance

Acknowledgements

This version benefited from comments from DANE, including Mara Bravo, Julieth Solano, and Arleth Sorith. Further comments and feedback will be incorporated before the document is finalized. This version benefited from significant contributions from Andrés Clavijo, Lead Researcher and Coordinator for Colombia, Data-Pop Alliance; Natalie Shoup, Program Manager, Data-Pop Alliance; Carson Martinez, Research Assistant, Data-Pop Alliance; and Lauren Barrett, Media and Communications Strategist, Data-Pop Alliance.

Funding

Funding for this paper was provided by the World Bank Group whose support is gratefully acknowledged, as well as the Rockefeller Foundation as part of their core support to Data-Pop Alliance's activities.

Disclaimer

The views presented in this paper are those of the author and do not represent those of his institutions.

Suggested Citation

"Opportunities and Requirements for Leveraging Big Data for Official Statistics and the Sustainable Development Goals in Latin America." Data-Pop Alliance (Harvard Humanitarian Initiative, MIT Media Lab and Overseas Development Institute). May 2016.

Introduction

National Statistical Offices (NSOs) remain a pillar of democratic societies, but increasingly contest with new data producers both from the public and the private spheres. New data sources—such as social media data, cell-phone data, satellite data, etc.—have created new opportunities and challenges for the production of statistics, their dissemination and engagement with beneficiaries; and leading to discussions about a new set of responsibilities that goes beyond pure measurement towards informing or even creating knowledge within societies.[1] Simultaneously, NSOs are getting prepared for a new task: the "Data Revolution"; this global development puts them at the center of the Post-2015 agenda, and their contribution in measuring the Sustainable Development Goals (SDGs) will inevitably be important.

There is certain excitement that Big Data could be one element to help NSOs fulfill their responsibility. The advent of Big Data will influence the business of organizations whose core business lies in the production of statistical data. Not surprisingly, the discussions on "Big Data and Official Statistics" originated within NSOs statistical systems are well-established. However, in developing countries, many NSOs are still struggling with basic operating challenges, such as access to administrative data, poor collaboration between different governmental agencies, poor financial resources and capacities, and the absence of legislative frameworks. These challenges question the extent to which NSOs might be capable to actively engage with the Big Data.

The rise of Big Data does not automatically imply that it will favor societal prosperity; the Snowden revelations, oppressive governments' use of data to identify and arrest innocent people,[2] and the increasing power of algorithms to enable discrimination against the underprivileged[3] give sufficient indication that Big Data can also harm democratic and human rights-based societies. The societal discourse about how a data-driven world should be shaped has only just begun; while—simultaneously—we will continue to create more and more data every day.

NSOs are governed by legal democratic frameworks and have the general tools and know-how to work with data in the most sensitive manner, under the premise of contributing to the wellbeing of societies in accordance with the first of the Fundamental Principles of Official Statistics. That is why they need to be key players in shaping the Big Data ecosystems of their respective countries and regions. In countries where they are recognized as trusted third parties, they will be crucial in the context of sharing data and forming a counterbalance to the interests of the private sector and governmental actors, including safeguarding privacy and the quality of the data.

Even from an opportunistic perspective, it would only be reasonable for NSOs to engage with Big Data as it becomes more important and as governments across the globe exercise influence in this field. If NSOs show leadership and become authorities on Big Data, they might receive

[1] Giovannini 2010.

[2] Culzac 2014.

[3] Tufekci 2014.

the recognition and prioritization from governments they so urgently need (and with that more resources). Big Data can be strategically important to NSOs in several other respects. Given their likely higher level experience in developing techniques and standards related to data collection, curation and release (for example, metadata and data anonymization), NSOs will have a clear role to play in issuing guidelines in these areas for their own statistical products and for other agencies in the national statistics system.

In Latin America[4], NSOs could clearly benefit from this opportunity. Compared to its counterparts in other developing regions, the Latin American statistical system is relatively strong, and the experiences of measuring the Millennium Development Goals (MDGs) in the last 15 years provide some well-established processes and tools. Still, the specific character of the SDGs presents new challenges, while many of the old ones remain unresolved: there are wide variations between the quality of NSOs within the region and even the more advanced NSOs still struggle with limited access to administrative records, blurry legal frameworks, and a lack of territorial and disaggregated data—data that will play a key role in measuring the SDGs.

However, with statistical systems far better than in many other parts of the world, wide penetration in mobile and Internet technology, vibrant debates on Internet governance, and an impressive Open Data movement, Latin America could become solid ground for good practices in Big Data. And as illustrated in this document, some NSOs in Latin America are increasingly tackling the task ahead. They are working on pilots and projects and are researching the potential of Big Data.

This report highlights the opportunities and challenges that Big Data presents to NSOs in the Latin American region in the context of the SDGs, identifying ongoing Big Data activities NSOs and non-NSO actors are currently undertaking and providing recommendations for NSOs in the region to play a role in evolution of official statistics and the SDGs in the regional Big Data ecosystem. More broadly it aims at answering the following questions:

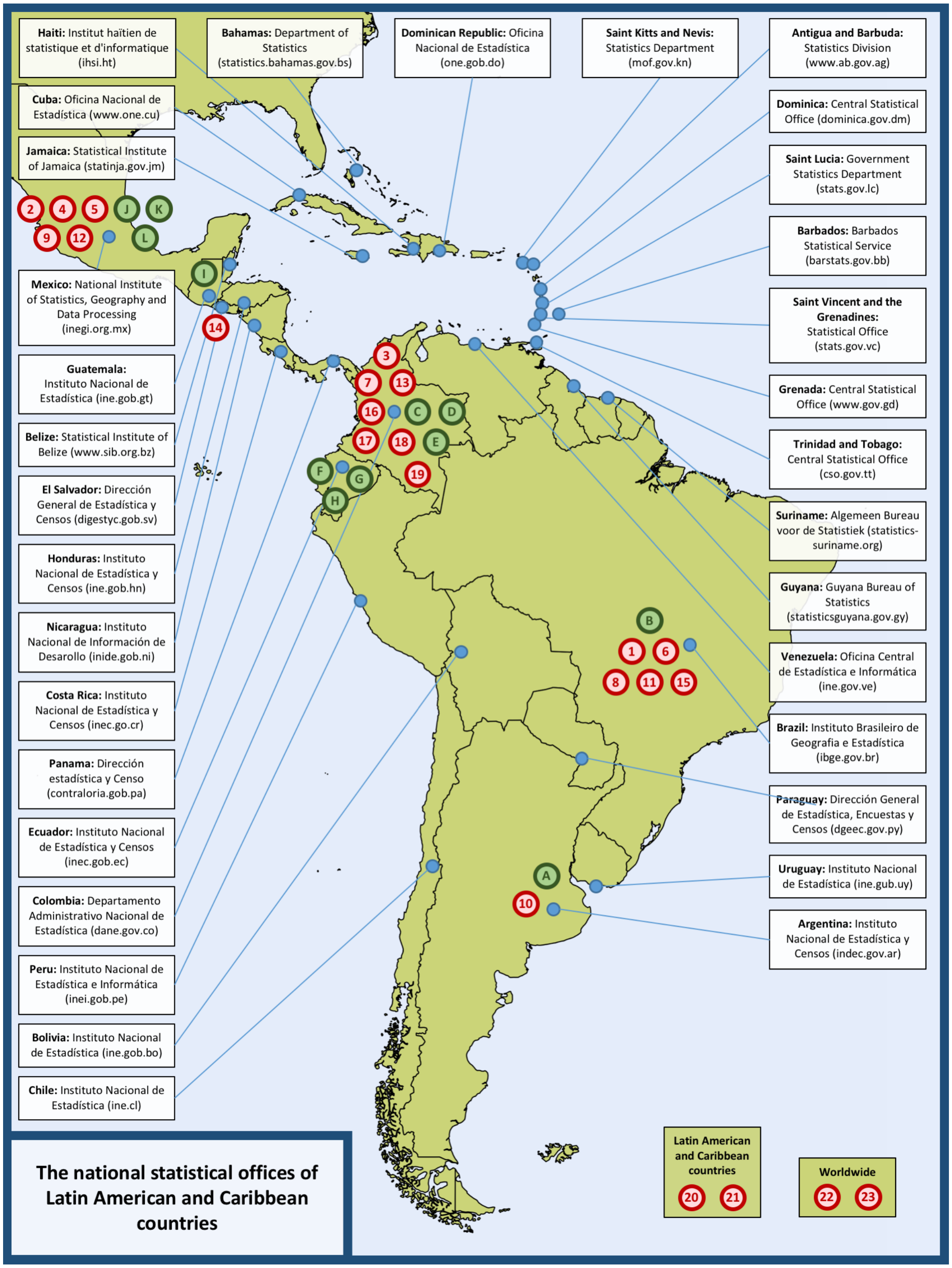
1. What is the current state of NSOs across Latin America?
2. How have and can NSOs engage with Big Data toward official statistics and the SDGs?
3. What kinds of new challenges do NSOs face in adopting Big Data?
4. How can these innovations be aligned towards national and regional strategies?

The rest of this paper is organized as follows. The first section of the paper describes the state of NSOs in Latin America, detailing their role in the context of the Post-2015 agenda, current challenges and conceptually the use of Big Data for official statistics and SDG measurement. In the next section of the paper, we look at the universe of Big Data activities that regional NSOs are currently undertaking, as well as the activities of other actors across the broader ecosystem of Big Data and statistical systems in Latin America. This includes an overview of the on-going

[4] In this paper, we refer to Latin America or the Latin American and Caribbean region including all countries and islands on the American continent south of the United States of America.

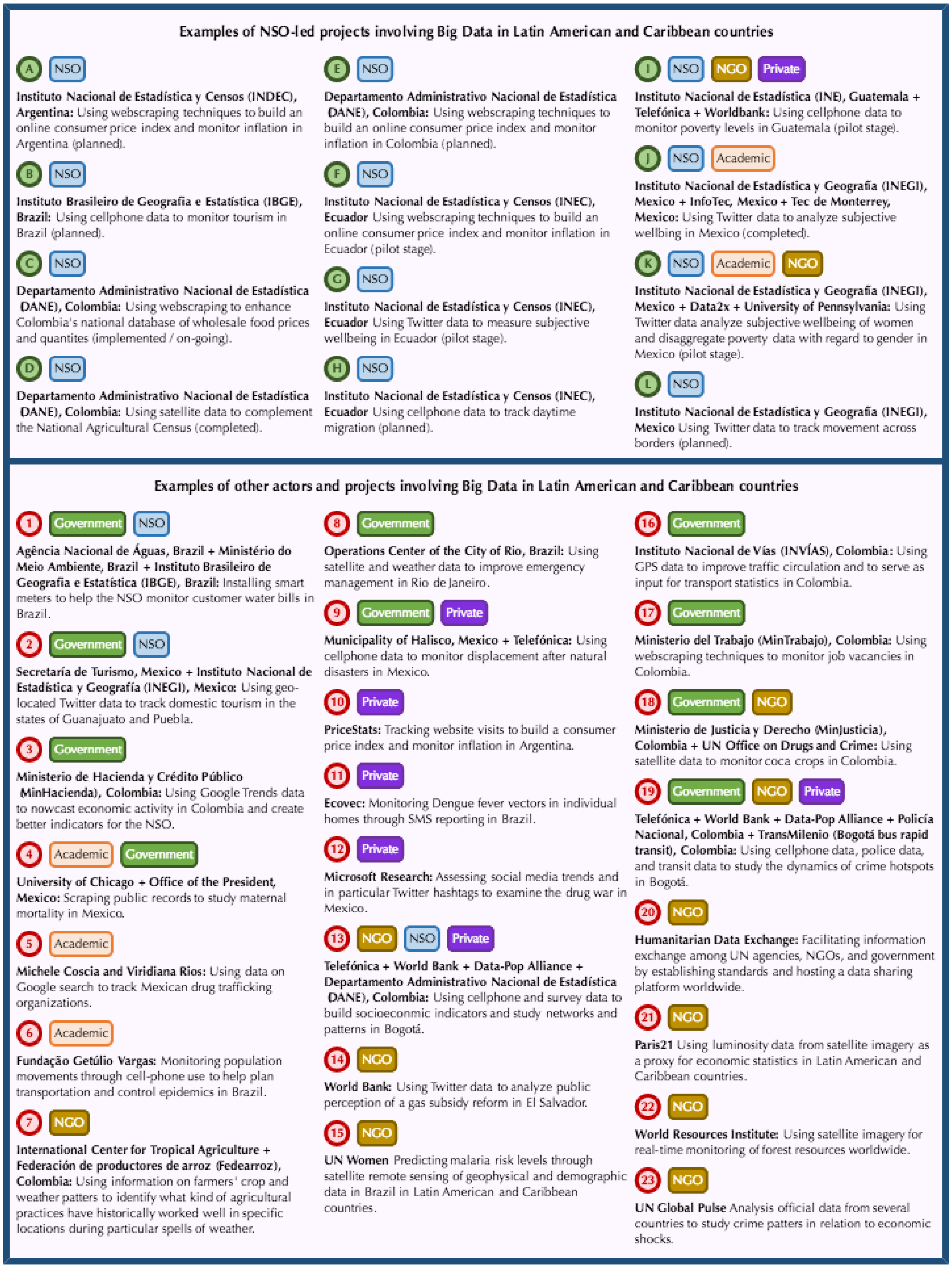
Big Data pilots and initiatives within and outside the statistical system. In the last sections of the paper, we analyze specific challenges for further NSO adoption of Big Data, provide recommendations towards next steps for NSOs to engage with Big Data and lastly discuss a series of recommendations for a regional roadmap for NSOs and other actors towards further regional engagement with Big Data.

Figure 1: Map of National Statistical Offices in Latin America.


Haiti: Institut haïtien de statistique et d'informatique (ihsi.ht)
Bahamas: Department of Statistics (statistics.bahamas.gov.bs)
Dominican Republic: Oficina Nacional de Estadística (one.gob.do)
Saint Kitts and Nevis: Statistics Department (mof.gov.kn)
Antigua and Barbuda: Statistics Division (www.ab.gov.ag)
Cuba: Oficina Nacional de Estadística (www.one.cu)
Dominica: Central Statistical Office (dominica.gov.dm)
Jamaica: Statistical Institute of Jamaica (statinja.gov.jm)
Saint Lucia: Government Statistics Department (stats.gov.lc)
2
4
5
J
K
9
12
L
Barbados: Barbados Statistical Service (barstats.gov.bb)
Mexico: National Institute of Statistics, Geography and Data Processing (inegi.org.mx)
I
14
Saint Vincent and the Grenadines: Statistical Office (stats.gov.vc)
Guatemala: Instituto Nacional de Estadística (ine.gob.gt)
3
7
13
Grenada: Central Statistical Office (www.gov.gd)
Belize: Statistical Institute of Belize (www.sib.org.bz)
16
C
D
17
18
E
Trinidad and Tobago: Central Statistical Office (cso.gov.tt)
El Salvador: Dirección General de Estadística y Censos (digestyc.gob.sv)
F
G
19
H
Suriname: Algemeen Bureau voor de Statistiek (statistics-suriname.org)
Honduras: Instituto Nacional de Estadística y Censos (ine.gob.hn)
Guyana: Guyana Bureau of Statistics (statisticsguyana.gov.gy)
Nicaragua: Instituto Nacional de Información de Desarollo (inide.gob.ni)
B
1
6
8
11
15
Venezuela: Oficina Central de Estadística e Informática (ine.gov.ve)
Costa Rica: Instituto Nacional de Estadística y Censos (inec.go.cr)
Brazil: Instituto Brasileiro de Geografia e Estadística (ibge.gov.br)
Panama: Dirección estadística y Censo (contraloria.gob.pa)
Paraguay: Dirección General de Estadística, Encuestas y Censos (dgeec.gov.py)
Ecuador: Instituto Nacional de Estadística y Censos (inec.gob.ec)
Uruguay: Instituto Nacional de Estadística (ine.gub.uy)
A
10
Colombia: Departamento Administrativo Nacional de Estadística (dane.gov.co)
Argentina: Instituto Nacional de Estadística y Censos (indec.gov.ar)
Peru: Instituto Nacional de Estadística e Informática (inei.gob.pe)
Bolivia: Instituto Nacional de Estadística (ine.gob.bo)
Chile: Instituto Nacional de Estadística (ine.cl)
The national statistical offices of Latin American and Caribbean countries
Latin American and Caribbean countries
20
21
Worldwide
22
23

Wikipedia and elaboration by Gabriel Pestre

Figure 2: Selected Big Data Projects in Latin America.



Examples of NSO-led projects involving Big Data in Latin American and Caribbean countries

A NSO

Instituto Nacional de Estadística y Censos (INDEC), Argentina: Using webscraping techniques to build an online consumer price index and monitor inflation in Argentina (planned).

B NSO

Instituto Brasileiro de Geografia e Estatística (IBGE), Brazil: Using cellphone data to monitor tourism in Brazil (planned).

C NSO

Departamento Administrativo Nacional de Estadística (DANE), Colombia: Using webscraping to enhance Colombia's national database of wholesale food prices and quantites (implemented / on-going).

D NSO

Departamento Administrativo Nacional de Estadística (DANE), Colombia: Using satellite data to complement the National Agricultural Census (completed).

E NSO

Departamento Administrativo Nacional de Estadística (DANE), Colombia: Using webscraping techniques to build an online consumer price index and monitor inflation in Colombia (planned).

F NSO

Instituto Nacional de Estadística y Censos (INEC), Ecuador Using webscraping techniques to build an online consumer price index and monitor inflation in Ecuador (pilot stage).

G NSO

Instituto Nacional de Estadística y Censos (INEC), Ecuador Using Twitter data to measure subjective wellbeing in Ecuador (pilot stage).

H NSO

Instituto Nacional de Estadística y Censos (INEC), Ecuador Using cellphone data to track daytime migration (planned).

I NSO NGO Private

Instituto Nacional de Estadistica (INE), Guatemala + Telefónica + Worldbank: Using cellphone data to monitor poverty levels in Guatemala (pilot stage).

J NSO Academic

Instituto Nacional de Estadística y Geografía (INEGI), Mexico + InfoTec, Mexico + Tec de Monterrey, Mexico: Using Twitter data to analyze subjective wellbing in Mexico (completed).

K NSO Academic NGO

Instituto Nacional de Estadística y Geografía (INEGI), Mexico + Data2x + University of Pennsylvania: Using Twitter data analyze subjective wellbeing of women and disaggregate poverty data with regard to gender in Mexico (pilot stage).

L NSO

Instituto Nacional de Estadística y Geografía (INEGI), Mexico Using Twitter data to track movement across borders (planned).

Examples of other actors and projects involving Big Data in Latin American and Caribbean countries

1 Government NSO

Agência Nacional de Águas, Brazil + Ministério do Meio Ambiente, Brazil + Instituto Brasileiro de Geografia e Estatística (IBGE), Brazil: Installing smart meters to help the NSO monitor customer water bills in Brazil.

2 Government NSO

Secretaría de Turismo, Mexico + Instituto Nacional de Estadística y Geografía (INEGI), Mexico: Using geo-located Twitter data to track domestic tourism in the states of Guanajuato and Puebla.

3 Government

Ministerio de Hacienda y Crédito Público (MinHacienda), Colombia: Using Google Trends data to nowcast economic activity in Colombia and create better indicators for the NSO.

4 Academic Government

University of Chicago + Office of the President, Mexico: Scraping public records to study maternal mortality in Mexico.

5 Academic

Michele Coscia and Viridiana Rios: Using data on Google search to track Mexican drug trafficking organizations.

6 Academic

Fundação Getúlio Vargas: Monitoring population movements through cell-phone use to help plan transportation and control epidemics in Brazil.

7 NGO

International Center for Tropical Agriculture + Federación de productores de arroz (Fedearroz), Colombia: Using information on farmers' crop and weather patters to identify what kind of agricultural practices have historically worked well in specific locations during particular spells of weather.

8 Government

Operations Center of the City of Rio, Brazil: Using satellite and weather data to improve emergency management in Rio de Janeiro.

9 Government Private

Municipality of Halisco, Mexico + Telefónica: Using cellphone data to monitor displacement after natural disasters in Mexico.

10 Private

PriceStats: Tracking website visits to build a consumer price index and monitor inflation in Argentina.

11 Private

Ecovec: Monitoring Dengue fever vectors in individual homes through SMS reporting in Brazil.

12 Private

Microsoft Research: Assessing social media trends and in particular Twitter hashtags to examine the drug war in Mexico.

13 NGO NSO Private

Telefónica + World Bank + Data-Pop Alliance + Departamento Administrativo Nacional de Estadística (DANE), Colombia: Using cellphone and survey data to build socioeconmic indicators and study networks and patterns in Bogotá.

14 NGO

World Bank: Using Twitter data to analyze public perception of a gas subsidy reform in El Salvador.

15 NGO

UN Women Predicting malaria risk levels through satellite remote sensing of geophysical and demographic data in Brazil in Latin American and Caribbean countries.

16 Government

Instituto Nacional de Vías (INVÍAS), Colombia: Using GPS data to improve traffic circulation and to serve as input for transport statistics in Colombia.

17 Government

Ministerio del Trabajo (MinTrabajo), Colombia: Using webscraping techniques to monitor job vacancies in Colombia.

18 Government NGO

Ministerio de Justicia y Derecho (MinJusticia), Colombia + UN Office on Drugs and Crime: Using satellite data to monitor coca crops in Colombia.

19 Government NGO Private

Telefónica + World Bank + Data-Pop Alliance + Policía Nacional, Colombia + TransMilenio (Bogotá bus rapid transit), Colombia: Using cellphone data, police data, and transit data to study the dynamics of crime hotspots in Bogotá.

20 NGO

Humanitarian Data Exchange: Facilitating information exchange among UN agencies, NGOs, and government by establishing standards and hosting a data sharing platform worldwide.

21 NGO

Paris21 Using luminosity data from satellite imagery as a proxy for economic statistics in Latin American and Caribbean countries.

22 NGO

World Resources Institute: Using satellite imagery for real-time monitoring of forest resources worldwide.

23 NGO

UN Global Pulse Analysis official data from several countries to study crime patters in relation to economic shocks.

Wikipedia and elaboration by Gabriel Pestre

1 The State of Latin American NSOs: Overall Context and Concepts

1.1 The Role of National Statistical Offices in Latin America and the Caribbean

Countries in Latin America—with significant geographical and socio-economic disparities and therefore, a diverse array of regional statistical challenges—has developed a strong tradition in official statistics, centered around their National Statistical Offices (NSOs). As noted in the 2010 Report from the Economic Commission for Latin America and the Caribbean (ECLAC), these NSOs by law govern data collection for the production and dissemination of statistics, manage strategy for long-term national survey implementation, and typically provide guidelines and leadership within largely decentralized national statistical systems.[5]

NSOs in the region have made considerable advances in the last decade in official data collection, production and dissemination in the following areas: population censuses, household surveys, income and expenditure surveys, national accounts and economic statistics, price statistics, gender statistics, vital statistics, education statistics, environmental statistics, and ICT statistics. The region has a great tradition in conducting censuses leading to more or less solid ground truth data across the region. Almost all Latin American and Caribbean countries have conducted a population census in the last ten years, and around half of them conduct household surveys every five years.[6]

As NSOs continue to develop and meet statistical challenges in the region, the post-2015 development agenda and the creation of the Sustainable Development Goals (SDGs) have brought attention to and highlighted the need for NSOs to address ongoing statistical challenges as well as incorporate innovative approaches and opportunities through new data sources. The adoption of the SDGs involves a complex set of goals with 169 targets covering environmental, economic, social, and governance dimensions. The first draft includes 310 indicators aligned to the targets. Experience with the MDGs have taught us that new measurements are needed beyond national averages and aggregation; the SDGs aim to accurately identify the most vulnerable, marginalised, and poorest people, requiring data at a local level, disaggregated by demographic groups (such as income, gender, age, race, ethnicity, migratory status, disability, geographic location, and other characteristics relevant in national contexts). However, this level of disaggregated data is currently not available in many countries. For some of the indicators, appropriate data is not even available in aggregated form.

The development of the post-2015 development agenda places NSOs at the core of the SDG activities. The UN Secretary-General's Independent Expert Advisory Group on a Data Revolution for Sustainable Development (IEAG) in their report "A World that Counts"

[5] Economic Commission for Latin America and the Caribbean (ECLAC/CEPAL) 2010.

[6] Economic Commission for Latin America and the Caribbean (ECLAC/CEPAL) 2010.

demand that UN member countries strengthen their NSOs' capacities to accomplish a "data revolution." There are several reasons why NSOs must have an active role in the production and collection of data for the SDGs:

1. NSOs typically have more experience than other actors in gathering data.
2. Data on development and society is a public good; therefore, it makes sense that public bodies produce it and that we build their capacity to do so.
3. NSOs typically retain the highest methodological standards.
4. Data about a country should ideally be produced and owned by that country to promote allocative efficiencies, thereby increasing legitimacy and use by policymakers.
5. States will play a central role in driving national-level progress towards meeting the SDGs. Officials require data to guide their policymaking, and official bodies should be responsible for gathering it.

Further, it should be noted here that, independent of the specific potential for the measurement of the SDGs, NSOs need to be involved in the discussion in any case. It is their mandate to foster knowledge about and among the societies that mandated them to do so. As stated by the Principle 1 of the Fundamental Principles of Official Statistics, "Official statistics provide an indispensable element in the information system of a democratic society, serving the Government, the economy and the public with data about the economic, demographic, social and environmental situation." If we as the international community believe in the democratic necessity of official statistics, we see that this is a political question of why NSOs must engage with Big Data, not simply a technical question of whether or not and how they should 'use' big data streams.[7]

Big data must, gradually, in time, become part of the resources and tools leveraged to fulfill this mandate and to provide a picture of a country, its economy and its population that can be turned into knowledge. There is a risk that those who will report on the state of societies using Big Data will eventually hold a great deal of power created by knowledge produced outside of the scope of democratic decisions and oversight. With their specific mandate, NSOs could play the role of gatekeepers to ensure the quality of new data sources but also to manage the downsides of the data revolution, such as privacy and confidentiality issues, as they are guided by established legal frameworks.[8]

However, NSOs are no longer the only actors producing and collecting data on society. As a result of digitalization and the on-going increase of web-data, a growing number of new actors have become producers of data. For example, market researchers are gaining accurate insights about their customers (and thus about citizens) through the automated analysis of digital datasets in high speed processing. Data that is generated passively by humans and machines in

[7] Letouzé 2013.

[8] Letouzé 2013.

high volume and with high velocity, such as data from social media or mobile phone records, is termed “big data.” Additionally, digital technologies have lowered the costs of producing and publishing data; they have eased the distribution and visualization of data and have hence democratized access to data and create new use cases for it. In Latin America this can be seen prominently in countries like Uruguay and Brazil that have actively embraced the Open Data movement.[9]

The intelligent mediation of data today becomes an essential element for getting a rich idea of societies and citizens’ demands, and hence evidence-based policymaking. These developments sparked a vital discussion about the role of NSOs and the need of NSOs to evolve from pure data producers to facilitators of comprehensible information that can be turned into knowledge about reality. This role encompasses all stages of the statistical process, from data collection to dissemination.[10] The IEAG report highlights the need for an institutional change towards innovation and much more effective use of technology to improve the performance of all actors involved in the production and collection of data.[11]

1.2 The State of NSOs in LAC: Ongoing Challenges

As the role of NSOs continues to evolve, NSOs in LAC must consider three categories of challenges that currently hinder official statistical activities across the region: data quality overall, coverage, and legislative considerations.

Quality, including *reliability*, *timeliness* (i.e. the length of time between the reference period and when statistics are made available), *interpretability* (or availability of metadata, that reflects the ease with which the user may understand and properly use the data) and *compliance* (i.e. the extent to which the statistics comply with the relevant international standards. For example, while the number of censuses increases, data quality does not always improve. As has been demonstrated in the 2010 round of censuses, Paraguay and Chile had several problems with their last census, “with a sub enumeration estimated at around 26.0% and 9.3%, respectively. Those figures for census omission, after decades of experience in data collection, are inconceivable.”[12] Even in countries with strong statistical systems, we see a number of problems. The Colombian census was planned to be conducted in 2015 and has been postponed and is now scheduled to take place in 2016. The agricultural census was not renewed for more than 40 years, until it was successfully conducted in 2014. In Brazil, the 2015 population count was recently canceled, even though it had been planned for years.[13] Most often budget cuts (for example for some countries due to the recent decrease of petrol prices) and poor long-term planning are the reasons for these fallouts.

Poor survey design sometimes lead to lack of qualitative data sets: For example, in Bolivia, in its last census a high percentage of women replied with “Unknown” when asked if they have had

[9] Open Data Research Network 2014.

[10] Giovannini 2010.

[11] Data Revolution for Sustainable Development (IEAG) 2014.

[12] Cavenaghi 2015.

[13] Cavenaghi 2015.

children. As a result, it is unknown whether half the population over the age of 15 already gave birth.[14] A pilot project on maternal mortality carried out by CEPAL/CELADE identified the difficulties in estimating maternal mortality in Latin America because of the lack of certification or registry in areas inhabited by indigenous populations or in remote zones.[15]

Vital statistics and civil registration systems—which will be particularly important for measuring the SDGs—are often weak across the region. A large proportion of LAC regions lack data on such variables as the age of mothers, the birth weight of children, and the place of residence or socio-economic characteristics of the parents. Data on causes of death is also frequently imprecise or non-existent, preventing understanding of the true levels of risk and prevalence of disease in the countries, as well as hindering the formulation of epidemiological mortality profiles.[16] In contrast to their expertise on surveys and censuses, many NSOs still struggle with accessing and using administrative data—as described below—although improved access has been achieved in the last couple of years.[17]

Coverage refers to the extent to which the statistics meet the requirements in terms of variables, detail, frequency, measurement units, historical coverage and availability. Poverty data, the quantification of inequality measurements, and the disaggregation of information for the identification of social, economic and environmental gaps also remain problematic.[18] In other areas, for example in terms of gender indicators, the current challenge goes beyond the disaggregation of indicators for monitoring the post-2015 development agenda. They point to the need for more active interaction between technicians, who design and use the information, and thematic specialists, for example on gender.[19]

Similar to many other developing countries data in LAC is often insufficiently disaggregated at sub-national level, making it hard for policy makers or communities to compare their progress with that of other communities or the country as a whole.[20] This is particularly striking in Latin America where there are enormous gaps between socio-economic levels in rural compared to urban areas as well as between different groups such as indigenous and Afro-descendants. This creates hurdles to providing solid data for measuring the progress of the SDGs.

Legislative considerations: Many Latin American NSOs still lack adequate institutional and legal frameworks. This has implications on good business practices and transparency. Instead, many rely on non-mandatory or completely voluntary regulations, such as the National Code of Good Practices—although there are convincing arguments for creating autonomous and apolitical statistical bodies and conditions under which statistics are isolated from politics.[21] Also, budgetary management is most often not independent from the rest of the government. Additionally, in many LAC countries senior manager positions are still selected by ruling

[14] Cavenaghi 2015.

[15] Cobos, Miller, and Salguero 2013.

[16] Economic Commission for Latin America and the Caribbean (ECLAC/CEPAL) 2010.

[17] Economic Commission for Latin America and the Caribbean (ECLAC/CEPAL) 2010.

[18] Economic Commission for Latin America and the Caribbean (ECLAC/CEPAL) 2015a.

[19] Economic Commission for Latin America and the Caribbean (ECLAC/CEPAL) 2015a.

[20] Economic Commission for Latin America and the Caribbean (ECLAC/CEPAL) 2015a.

[21] Khan and Stuart 2015.

politicians and filled by either senior civil servants or government ministries.[22] Events like the scandal surrounding the National Institute of Statistics and Census of Argentina (INDEC), which had been allegedly manipulated by the Kirchner government, discredit NSOs across the region and call their trustworthiness into question.

Limited trust and transparency about statistical processes presents major impediments. Certainly, the adoption and revision of statistical legislation to ensure the independence of NSOs in many countries takes a crucial step towards improving their credibility, as has been seen in Mexico (see Box 1). Fortunately, we see a trend throughout the region towards the introduction of publicly managed systems in which senior management positions in statistical offices are filled through a competitive recruitment mechanism.[23]

Additionally, poor legislation leads to undefined mandates regarding data collection and access. For example the generation of statistics from administrative records is still limited, because there are often no clear laws allowing NSOs to request this data from other agencies. Inter-operationality between governmental institutions most often poses a challenge and many NSOs are competing rather than collaborating with other ministries and agencies. A positive development is that under Colombia's National Development Plan's article 150, DANE achieved better control of harnessing of administrative registers for statistical purpose.[24]

1.3 Defining "Big Data" for Official Statistics and the SDGs

Big data—as a new data source—is potentially interesting as an input for official statistics, for use both on its own and in combination with more traditional data sources such as sample surveys and administrative registers. It has the potential to produce more relevant and more timely statistics than traditional sources.[25] For instance, analyses of comments, search queries, or online posts can produce nearly the same results for statistical inference—more quickly and at lower cost—as household surveys and polls. Data on employment can be monitored for free and in real-time via Google Trends.[26]

This example further illustrates that there is more to the phenomenon than "big data" as a new source, just as there was more to industrialization than oil and electricity. With the advent of Big Data come along new actors, capacities, and instruments that are and will be shaped by society. When we speak of this wider phenomenon, whose transformative potential can be compared to that of industrialization, we speak of "Big Data" (in capital letters) and not just "big data."

Big Data refers not only to data but also to the institutions and wider ecosystem that produce and use it.[27] This ecosystem can be described as the union of Big Data "crumbs" (new kind of

[22] Economic Commission for Latin America and the Caribbean (ECLAC/CEPAL) 2010.

[23] Economic Commission for Latin America and the Caribbean (ECLAC/CEPAL) 2010.

[24] Congreso de la República de Colombia 2014.

[25] United Nations Statistical Commission 2014a.

[26] Hubbard 2011.

[27] Pentland 2012.

Box 1: INEGI as Best Practice in the Region

The National Institute of Statistics and Geography of Mexico (INEGI) clearly provides an example of best practices for the development of statistics in the region. In 2008 it has been transformed into a fully autonomous institution, governed by a board of directors appointed by the President of the Republic and the Senate. INEGI's operations are supported by a law that allows the institution to effectively exercise a leading role throughout the national statistical system. INEGI clearly benefits from solid financial resources allowing them to hire highly professionalized staff. INEGI is the technical counterpart for various national endeavors, such as the implementation of the Open Data strategy. Due to Mexico's membership to the OECD, INEGI also gains insights and expertise from institutionalized exchange on an international level.

passively generated data), "capacity" (as the technical and human capacity to yield insights from this data) and "community" (new actors from the private sector and the research community for example).[28]

Big Data has three major characteristics and implications that highlight its potential for complementing and augmenting the existing work of NSOs (see Box 1).

1. Big Data provides new sources of data

First of all, it sometimes remains blurry what kind of data can actually be defined as "big data". Currently, we witness some ambiguity in the use of terms such as open data, "smart data," "thick data," big data, and Big Data (capitalized). All of these will be important components of realizing a "data revolution." But big data have highly distinct qualities that differentiate them from conventional sources of data; they are of large volume and can be composed of all kinds of data-generating sources and hence be both structured and unstructured. For instance, while administrative records (one of the main sources used by many NSOs) pose large amounts of data and big spreadsheets, they would not be considered Big Data until the velocity increases, for example if administrative data is collected on a daily basis.[29] And while establishing a data warehouse is an important step towards processing Big Data sets, its main feature is to store large sets of structured data, which often constitute big data but not necessarily Big Data.

2. Big Data provides greater diversity of data sources

This leads to the second issue: Big Data is not about the data nor about its size, as has been remarked by various scholars.[30] It is "*different* data that may contain signals unavailable only

[28] Pentland 2012.

[29] United Nations Economic Commission for Europe (UNECE) 2013.

[30] King 2013.

a few years ago that 'we' are yet to know how to read and use,"[31] and which is not actively and intentionally requested by statisticians or researchers. In contrast to data from traditional sources collected with the goal of answering a question, Big Data might give answers to questions that have not even been asked. It is new data and should be considered "as digital traces of human actions passively generated by individuals."[32]

3. Big Data has the potential to complement and improve ongoing statistical activities through its four functions

Big Data as an ecosystem has the potential to improve and complement official statistic activities by replacing particular indicators and measurement processes. Big Data can feed into the statistical process through its four functions:

1. **Descriptive**—via maps, descriptive statistics, visualizations, etc;
2. **Predictive**—to make inferences about current conditions and forecasts about future events;
   (a) Predicting as *proxying*—where Big Data is used to predict the concomitant level of another variable—poverty for instance; this is also referred to as *inference* or *nowcasting*.
   (b) Forecasting, where the likelihood of some events in the near or distant future is assessed.
3. **Prescriptive**—also referred to as diagnostic—to draw causal inferences with Big Data, where CDR analytics will help unveil casual relationships linking cell phone usage to outcome, or more generally help prescribe specific interventions.
4. **Discursive**—also referred to as engagement—which "concerns spurring and shaping dialogue within and between communities and with key stakeholders", recognizing that "the longer-term potential of Big Data lies in its capacity to raise citizens' awareness and empower them to take action."

Big Data experimentations can be applied to processes, outcomes, and related SDGS that are:[33]

- Correlated with (i.e. show in) trends and patterns in data production of some kind;
- Currently monitored through traditional means (providing 'ground truth' data without which calibration is impossible, or requires making assumptions);
- Deemed relatively more 'important' in universal terms (e.g. income poverty, health and education outcomes) as well as in contextual terms;

[31] Letouzé 2013.
[32] Letouzé 2013.
[33] Letouzé 2015.

- Applicable to 'new' kinds of sectors and goals, such as social cohesion, crime prediction or subjective well-being.

Some argue that these instruments will be far cheaper than traditional data collection, particularly surveys, which are still highly expensive and in some countries cannot yet be done electronically. Additionally, Big Data could contribute to improving some aspects of the quality of statistics, such as timeliness and completeness, without compromising their relevance, impartiality, and methodological soundness.[34] It could also complement or replace other more traditional ways of measuring facets of human reality, such as mortality, violence, or hunger, as various initial research projects (some of them documented in this paper) have shown.[35]

Big Data could also help fill data gaps in thematic areas and monitor goals where data is scarce; this is particularly salient in the context of the SDGs. The overall goal of the Post-2015 agenda is to eliminate global poverty as stated in Goal 1 (End Poverty) of the SDGs, reflected in Goal 10 (Reduce inequality), and in various indicators of other goals. Poverty data—mainly collected via expensive household surveys—however is scarce in many countries, particularly on a disaggregated (i.e. when representing small geographic units, such as cities, towns and villages) and up-to-date level. Big Data offers an opportunity to close this gap. In developed regions, there has been research conducted using social media in order to measure socio-economic levels. However, these data sources already pose demographic biases in developed regions, which are even bigger in the Global South. Mobile phones at the same time usually have high penetration and hence offer more representative data, although even in this case, representativeness is not guaranteed.[36]

Foremost, Big Data defines a turning point in the production of official statistics and the creative, relevant, and responsible combination of such statistics with non-official statistics. If implemented, it will dismantle the traditional paradigm of statistical systems at all levels of implementation and induce an institutional shift. Big Data will affect NSOs' work on various levels, including data collection, the management of data quality, data aggregation, data analysis (or service production) and lastly, data visualization and allocation.

The table in Annex 2 highlights and references uses of Big Data toward monitoring the SDGs.

In addition to innovation in collection and use of current resources, the data revolution also points to the possibility of Big Data for measuring the SDGs, and the role NSOs may play in engaging these resources. As the conversations at the global level point to the possibilities of leveraging Big Data for statistics, how have and can Latin American NSOs participate in this big data revolution? What unique challenges do they face?

[34] United Nations Statistical Commission 2014a.

[35] Letouzé 2015.

[36] A penetration rate of 100 or more does not mean that there hundred percent of a population actually own and use a phone.

Box 2: Big Data vs. big data

‘Big Data’ (capitalized) is this document (and others) refers to the ecosystem created by the concomitant emergence of “the 3 Cs of Big Data”:

- The 1st C stands for digital bread Crumbs—these pieces of data passively emitted and collected as by-products of people’s interactions with and uses of digital devices that provide unique insights about their behaviors and beliefs;
- The 2nd C stands of Big Data Capacities—what has also been referred to Big Data Analytics, that is the set of tools and methods, hardware and software, know-how and skills, necessary to process and analyze these new kinds of data—including visualization techniques, statistical machine learning and algorithms, etc.;
- The 3rd C stands for Big Data Communities, which describe the various actors involved in the Big Data ecosystem, from the generators of data to their analysts and end-users—i.e. potentially the whole population.

This ecosystem can be described and analyzed as a complex system, i.e. one where feedback loops exist between its different parts. At the most basic level new companies (e.g. Twitter or its future competitor) help generate new kinds of data that in turn lead to the development of new kinds of analytical tools, leading to new kinds of data, then new actors taking advantage of these new data and tools. It is possible that this new ecosystem may turn into or be part of a larger social phenomenon.

In contrast, big data refers to the 1st C above, i.e. the streams and sets resulting from humans leaving digital traces when using cell-phones (call detail records), credit cards (transactions), transportation (subway or bus records, EZ pass logs), social media and search engines, or having their actions picked up by sensors, whether physical (electrical meters, weigh sensors on a truck) or remote (satellites, cameras).

Box 3: Difference between Big Data and Open Data

Although Big Data and Open Data both typically take the form of large datasets being put to overlapping uses with similar tools, they are distinct concepts. As referred to above, Big Data can be characterized as an ecosystem of data generated about and by people as a by-produce of their use of digital devices and platforms (crumbs), the new tools and methods developed to collect, process and analyze this data (capacities), and the set of individuals and institutional actors that make use of the data and capacities (communities). The term Open Data generally refers to data that is made publicly accessible for everyone?s use, with a few legal and technical barriers as possible, this can include government data such as budget data, weather data or administrative records, scientific data as well as data hold by NGOs or private companies. In most cases however it contains structured data.

Many of the tools and capacities being developed and used with these data are common to both Open and Big Data. While it was once too costly and technically challenging to collect by-product information (in the case of Big Data) or widely distribute existing data (in the case of Open Data), the decreasing cost of storage and increasing capabilities of affordable processors and devices have made if possible for both Big Data and Open Data to develop.

Thus, while the new speed and scale with which it is now possible to store and process information has enabled both Big Data and Open Data to become mainstream concurrently, they are in fact different concepts: the former relates fundamentally to where data comes from and the later relates more to what is done with it.[a]

But indeed, data can be both Big and Open, as is the case with public databases of weather data collected through remote sensing, for example. While most data could theoretically be made Open, this is not always desirable for a variety of legal, ethical, technical, and/or financial reasons. For example, while the public sector and academic community might make interesting use of CDR datasets if they were made public, they are currently in the hands of cellphone providers, who have a financial disincentive from make this information available to their competitors and legal/ethical obligations towards their customers to keep them private. Conversely, there are some very interesting sources of data, such as records of consumer complaints against businesses, which can have benefits to society if they are made open, but don't happen to be Big Data (because they are actively reported by customers rather than collected passively through other uses). As part of the larger data ecosystem, Open Data can inform and improve other data analysis, for example in the context of Big Data. The Open Data Institute in London also speaks of the data spectrum to differentiate between different data sources and terms used in this context.[b]

[a] Gurin 2014.

[b] Open Data Institute 2015.

2 Engaging, Innovating, and Discovering Big Data in Latin America

2.1 Setting the Stage: the Burgeoning Ecosystem of Big Data

Like the rest of the world, Latin America is undergoing a digital revolution with increasing use and access to mobile technology and Internet connection. Similar to Africa and Asia, mobile technology has rapidly picked up in the last decade. New data sources, such as CDR data, generated by digital technologies—and defined as big data—are the fuel of the Big Data ecosystem. These kinds of data sources can be used to improve and complement statistical processes. However, their appropriateness for statistical operations greatly depends on availability. Notably data from mobile technology, social media data, and Internet data can only be relevant for statistical purposes if penetration rates are high enough. And not surprisingly, not all data sources are equally available in all countries. To estimate the potential of Big Data for the LAC region it therefore first needs to be assessed which kinds of data sources are actually available. One of the unique characteristics of Latin America is that its infrastructure is largely heterogeneous—this means that when evaluating data sources, or more specifically, a digitization index that takes into consideration Internet penetration, both the amount of and socio-economic status of the Internet user varies considerably.[37] Data and methodological bias and representation are elaborated on in section 3.5.

Internet

Uruguay, Chile, Costa Rica, and Argentina all have high Internet penetration rates as well as less inequality in terms of access (i.e. rural and urban divide and socio-economic levels).[38] This is not the case in other Latin American countries. While it is difficult to disaggregate numbers on current ICT data, old data from the Observatory for the Information Society in Latin America and the Caribbean (OSILAC)'s survey in 2010 shows that Internet access for wealthier households across the region exceeds the rates of access by the poorest segments by a factor of 44: "Indeed, there are strong correlations between Internet access and wider patterns of poverty, inequality, socio-economic class, and urbanization."[39] In rural areas many people still do not have access to the Internet at all. In Brazil and Colombia, the access gap between urban and rural households with fixed Internet connection exceeds 30 percentage points.[40] Industrial Internet use also sees a divide at the country level.[41]

[37] Katz 2015.

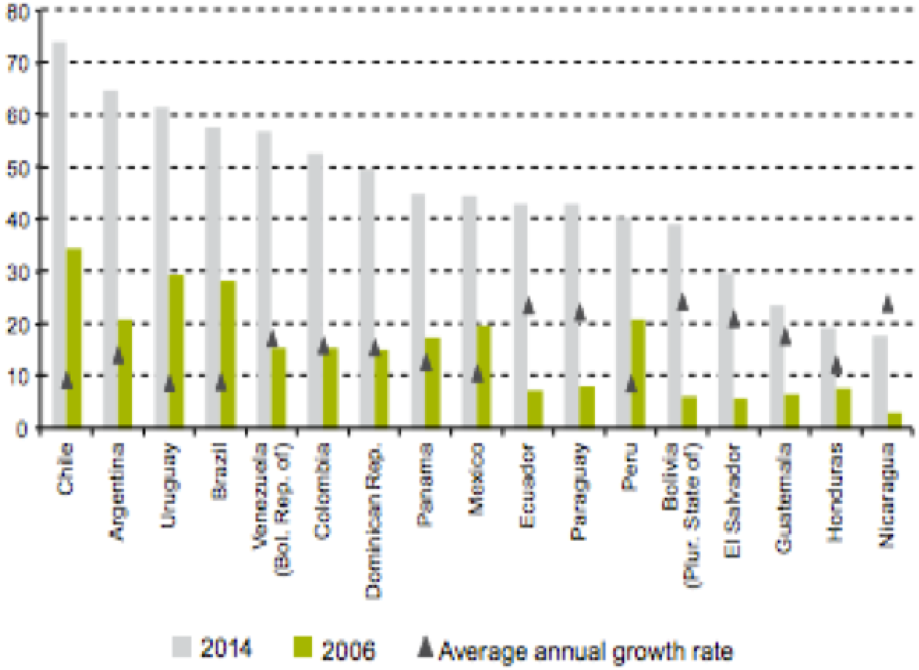
[38] Economic Commission for Latin America and the Caribbean (ECLAC/CEPAL) 2015b.

[39] Informa 2011.

[40] Economic Commission for Latin America and the Caribbean (ECLAC/CEPAL) 2015b.

[41] In commercial Internet use, for example, establishments in the manufacturing sector use the Internet to gather official information. The top three countries of highest percentage use for this are 70.5% in Argentina, 62.9% in Brazil, and 59.5% in Uruguay. (Katz 2015)

Figure 3: Internet Use by Percentage of Population, 2006 and 2014



Economic Commission for Latin America and the Caribbean (ECLAC/CEPAL). *The new digital revolution: From the consumer Internet to the industrial Internet.* 2015. URL: `http://repositorio.cepal.org/bitstream/handle/11362/38767/S1500587_en.pdf`

International Telecommunication Union. *World Telecommunication/ICT Indicators database, 19th Edition*. 2015. URL: `http://www.itu.int/en/ITU-D/Statistics/Pages/publications/wtid.aspx`

At the same time that we see a trend in growing gaps between Latin American countries' access to the Internet (see Figure 3), Central America exhibits the lowest penetration rates overall. Costa Rica and Nicaragua, with the greatest and least penetration rates respectively, are highlighted in Table 1 below, with Costa Rica clearly being a Central American exception. Overall Internet penetration rates are at 49.9% across the region.[42]

Table 1: 2013 Internet Usage and Population Statistics for Selected Countries in LAC

Country	Population (2014 Est.)	Internet Usage (31 Dec. 2013)	% Population (Penetration)
Argentina	43 024 374	32 268 280	75,0
Bolivia	10 631 486	4 199 437	39,5
Brazil	202 656 788	109 773 650	54,2
Chile	17 363 894	11 546 990	66,5
Colombia	46 254 297	28 475 560	61,6
Costa Rica	4 755 234	4 028 302	84,7
Guatemala	14 647 083	2 885 475	18,6
Mexico	120 286 655	59 200 000	49,2
Nicaragua	5 848 641	906 539	15,5
Ecuador	15 654 411	6 316 555	40,4
Panama	3 608 431	1 899 892	51,7
Paraguay	6 703 860	2 473 724	36,9
Peru	30 147 935	11 817 991	39,2
Uruguay	3 332 972	1 936 457	58,1

Internet World Stats. *Latin American Internet and Users and Population Statistics*. 2013. URL: `http://www.internetworldstats.com/stats10.htm`

Mobile

The Latin American mobile market is now the fourth largest in the world. Brazil, Mexico, and Argentina encompass the biggest markets because of their large populations and high penetration

[42] Internet World Stats 2013.

rates.[43][44] While the overall mobile penetration rate in Latin America rate is still far beyond 100 percent, a little over half of the population in the region is actually subscribed to a mobile service. However, this figure is expected to reach 60% by 2020, broadly in line with the global average.[45] In the same vein, overall global digitalization trends are also showing steady increases (see Table 1). Subscriber penetration rates range from a low of 37% in Mexico to a high of 77% in Costa Rica;[46] showing that there is not one single driver of the variation in penetration rates, and thus, differences in GDP per capita play a limited role. In comparison to other developing regions, mobile money services, which could also provide interesting data sources, have not yet taken off.[47]

Mobile networks and services are increasingly becoming the main method of accessing the Internet across Latin America. In 2011, the number of mobile broadband connections surpassed the number of fixed broadband connections.[48] Thanks to the increasing availability of lower cost models, smartphone ownership is increasing rapidly. For those who use mobile devices to access the Internet, there were 216 million individuals in September 2014, equivalent to an overall penetration rate of around 35%—already higher than the previous year's ownership statistic. By 2020, accessing Internet via mobile is forecasted to be just below 50% of the population.[49] An increase in both competition and innovation has allowed for more of both affordable smartphones and Internet access throughout the region.

The massive adoption of new information and communication technologies (ICT) has made possible greater generation (of digital data), communication, and dissemination of Big Data.

Social Media

The region is becoming one of the largest producers and consumers of social media, most notably on Facebook and Twitter [50] —producing a high amount of data that could be used for statistical

[43] Presently, América Móvil acting as a local operator (via its subsidiaries Claro and Telcel), Telef'onica (via Movistar), and Millicom (via Tigo) dominate the Latin American market. In Brazil, local operators Oi and Vivo have a significant share of the market. Interestingly enough, the top four Internet sites in every country in Latin America are of international origin—Google, Facebook, Microsoft, and Yahoo—with the exception of in Brazil (UOL) and Venezuela (Mercado Libre).

[44] Katz 2015.

[45] Mocanu et al. 2013.

[46] GSMA Intelligence n.d.

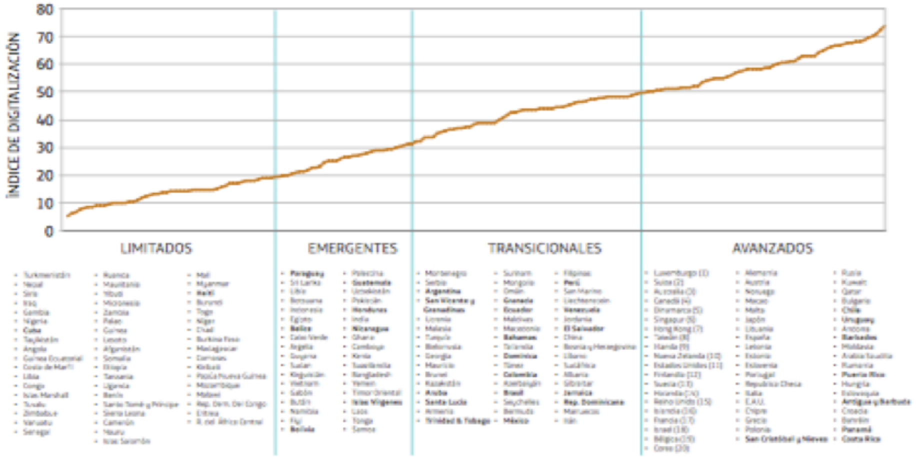
[47] "Millicom's Tigo Money is one of the only successful operator-led mobile money services that is active in five Latin American markets (Bolivia, El Salvador, Guatemala, Honduras, and Paraguay). In Bolivia, Tigo Money is responsible for money flows of nearly US$4 million per month and has around 700,000 customers. In Peru, Movistar recently launched a mobile money service in partnership with Mastercard, reaching a potential 16 million customers."

[48] GSMA Intelligence n.d.

[49] GSMA Intelligence n.d.

[50] The success of social media can also be also explained by its importance for political discussions and citizen engagement in many Latin American countries. In Brazil, social media was the main channel for debating the 2012 municipal elections and recent corruption trials, as well as organizing the protests surrounding the 2014 World Cup. In Mexico, Twitter has helped to spread social movements, such as the #YoSoy132 movement that emerged during

Figure 4: World Development of Digitalization 2013



Raúl Katz. *El ecosistema y la economía digital en América Latina*. 2015. URL: `http://cet.la/blog/course/libro-el-ecosistema-y-la-economia-digital-en-america-latina/`

purposes in some countries. Seven Latin American countries are among the world's top thirty in terms of Facebook users, with Brazil—referred to as the "Social Media Capital of the Universe" by the Wall Street Journal[51] —having the highest number of active users in the region, and Chile showing the highest ratio of users per capita.[52] Additionally, half of Latin American smartphone users are engaged with Twitter.[53] With over 41 million users, Brazil ranks second in the world in terms of the number of Twitter accounts and fifth globally in terms of usage, and it is the second-largest producer of tweets in the world.[54] Mexico ranks seventh in the world by number of Twitter accounts and has an estimated 11.7 million active users.[55]

2.2 NSOs and Big Data: Trends in Latin America

An increasing number of NSOs in the LAC region are showing interest in engaging with Big Data. International conferences such as the yearly World Statistics Congress of the International Statistical Institute and the UN and World Bank-led International Conference on Big Data for Official Statistics (the second edition of which took place in Abu Dhabi in October 2015) are

the 2012 presidential election. It has also become a tool for citizen journalists using it for the safe and anonymous publication of information about organized crime and the Drug War.

[51] Téllez 2015.

[52] Bibolini and Lancaster 2014.

[53] Reader 2015.

[54] Glickhouse 2013.

[55] Glickhouse 2013.

driving interest from stakeholders on a regional level as well. El Encuentro Mundial de Big Data took place in Bogotá in October 2015, as well as the Cartagena Data Festival in April 2015, also with partnership from DANE. The call for a data revolution and the demand for alternative and timelier measurement has undoubtedly sparked interest in Big Data approaches in Latin America, especially where the Post-2015 agenda is considered a political priority. Colombia and Costa Rica have weaved the goals into their national development plans. Hence, there are a number of ongoing pilots conducted by NSOs in the region, notably in Colombia, Mexico, and Ecuador. Throughout the region, the pilots range in Big Data use from web scraping and CDRs to social media, satellite, survey, and more; likewise the NSOs and their respective pilots range in what stage they are in—some are already planning pilots, like IBGE in Brazil, while others in Peru are still examining potential pilots.

Box 4: NSOs in Latin America: Colombia's DANE Moderno

In Colombia, DANE launched new high-level strategy called DANE Moderno which in 2014. DANE Moderno is expected to create a new mindset in the institution. This narrative follows the discourse of NSOs' responsibility to become knowledge producers in order to promote frameworks of democracy, such as informing citizens in ways that bolster citizen decision-making and in holding their governments accountable. The strategy also stresses the need for transparency with their citizens. "El DANE Moderno también significa un DANE de puertas abiertas, de respuestas amables y comprensibles, porque como lo he venido repitiendo, las cifras que producimos no son para quedarnos sentados en ellas, son para compartirlas con todos aquellos que las necesiten."[a][b] DANE Moderno has been referred to as an example of best practices by various international stakeholders (e.g. ODI and PARIS21). While the underlining premise of DANE Moderno is supposed to cultivate a new mindset and culture, it also has resulted in technical modernizations by further digitalizing technical processes, implementing new standards, and building a data warehouse. As part of DANE Moderno, DANE also launched a two-level innovation process in early 2015, which was inspired by the Innovation Lab at the Dutch Statistical Institute. This process encouraged employees to submit proposals on innovation, and employees submitted 84 ideas. Ten were selected and voted on by DANE employees via the intranet; four of these ten proposals had a Big Data component. An external jury eventually chose three final projects that are now ready for rollout.

[a] Roughly, "DANE Moderno means a DANE of open doors, of thoughtful and understandable responses, because the figures we produce are not to be held back, but to be shared with anyone that might need them."

[b] Cordero 2016.

As mentioned above, DANE in Colombia, INEGI in Mexico, and INEC in Ecuador are all taking a leading role in engaging with Big Data. However, this is expressed by very different approaches. In Colombia, Big Data is part of the above mentioned wider strategy—DANE Moderno—which is an innovative process to modernize the statistical operations on a structural and technical level in Colombia (Box 4). Big Data is regarded as one aspect of this process, while the overall technical changes, for example switching to Hadoop, favor this attempt. Big data activities are enforced at the directorial level and disseminated from there; a cross-department team was formed with support from external consultants and potential working areas for DANE have already been identified. Mexico and Ecuador, on the other hand, started with a technical and rather "hands-on approach" driven by internal champions. They initiated smaller pilots and played with data that was freely available via Twitter or Web scraping. Those on a management level seem to generally support their endeavors but they have not yet been given assignment from the top.

Not surprisingly, those countries that are part of the OECD (i.e. Mexico), or are aiming at becoming part of the OECD (i.e. Colombia and Peru in the early stages), or participate in other international working groups (i.e. those initiated by EuroStat and UNSD), are more progressive in their approach to Big Data. UNSD launched six working groups early 2015 that look at different aspects of Big Data. Mexico and Colombia participate in their activities and the Working Group on Big Data and the SDGs are jointly steered by the World Bank and Mecixo's INEGI. In particular the work of UNSD and UNECE's Sandox project has played a vital role for those Latin American NSOs that have had the privilege to participate. Pioneering projects, such as the work of the Dutch and the Estonian Statistical Offices, influence projects and pilots in the LAC region, making NSOs examine the feasibility of similar endeavors in their respective countries (see Annex 2).

In the Big Data Survey 2015 by the United Nations Global Working Group on Big Data for Official Statistics, three Latin American countries' NSO offices responded: Argentina, Ecuador, and Mexico. Among the Big Data projects reported, both Ecuador and Argentina were interested in creating real-time price indices from online data published by supermarkets and retail outlets.[56] There are a number of pilots planned in the region that correlate with this—the work of Alberto Cavallo at MIT, for one (see Annex 4) uses information from the Internet to generate price indices and measure Argentina's inflation,[57] and has been regarded positively by several NSOs. Some are already realizing or planning similar pilots (e.g. Ecuador, Argentina and Colombia). In Colombia, plans are in place to use the already existing data from SIPSA, the innovative national agricultural price information system (Annex 6), and compare and enrich it with web-scraped data. With its vast adoption by many NSOs around the world it is very likely that this approach in some way or another will be used to support the measurement of SDG 1 (Poverty Reduction).

Furthermore, freely available social media data has been used to conduct first pilots. Approaches that measure the well-being of citizens (similar to the pilot by the Dutch statistical office,[58] which has been presented at various events in the context of Big Data and statistics) (Annex 3) are garnering more and more international regard. INEGI in Mexico and INEC in Ecuador are currently working on similar pilots. Given Mexico's high numbers of estimated 11.7 million active Twitter users,[59] INEGI partnered with academic institutions[60] to develop the technical tools for measuring the subjective well-being through an analysis of tweets, and used

[56] Ecuador was also interested in creating a happiness index, driven by the Ministry of Well Being.

[57] Cavallo 2013.

[58] Daas and Loo 2013.

[59] Glickhouse 2013.

[60] Those include Infotec, CentroGeo, and CIMAT

tweets to monitor tourism movements (see Box 5). INEGI also plans to measure mental health among young women. In another pilot, INEGI used tweets to measure mobility and tourism movements during a long weekend in Puebla and Guanajuato. This was done in collaboration with the Mexican Secretary for Tourism.[61] Further research in this field is planned.

Box 5: Twitter for Tourism Monitoring in Mexico

In 2014, a working group on Big Data at INEGI conducted a pilot study to track domestic tourism from Twitter data, in order to contribute to the empirical modeling of individual tourist behavior. The objective of this pilot program was to identify the characteristics of an average Tweeting tourist in order to identify how many people travelled to Puebla and Guanajuato during the holiday weekend of February 1-3, 2014. The team of researchers from INEGI, in collaboration with the Mexican Ministry of Tourism, analyzed 60 million Tweets from January to July 2014, from the continuous 1% georeferenced sample that Twitter makes available for free. From this data, INEGI collected Tweets from the 7,955 Twitter users who Tweeted in Guanajuato (48%) and Puebla (52%) during the holiday. They then gathered all the Tweets sent by those users in the remainder of the target period (amounting to 827,424 total Tweets), and identified which users Tweeted from another state (presumably their homestate) after being in Guanajuato or Puebla, in order to map the origin of domestic tourism to those two areas during the holiday. The resulting estimates of domestic tourism to Guanajuato and Puebla were compared to estimates made by the respective offices of tourism of those two states.[a]

[a] Instituto Nacional de Estadística y Geografía (INEGI) de México 2015.

In contrast, we see fewer pilots and approaches from NSOs in CDR analysis. The exception is a pilot in Guatemala conducted by the World Bank and Telefónica Research as part of the World Bank's internal innovation contest, and involves the local NSO at a later stage. Some NSOs are interested in using CDRs for migration and tourist monitoring, as has been showcased by NSOs in Italy, Estonia, and the Netherlands. For example, IBGE is planning a pilot on monitoring tourist activities during the 2016 Olympics with CDR data. However, the lack of access to this data is currently hindering its preventing it from coming to fruition. Both Brazil and Colombia have reached out to mobile operators about accessing the data and are currently in discussions to sample data for pilots.

Similarly, there are already some pioneering examples in the application of satellite data. In Brazil, IGBE, with support from UN Women, are using remote sensing satellite data to predict both malaria risk and morbidity burden in pregnant women, especially along the Brazil-Guyana border (see Box 6). In Colombia, DANE used satellite data in a pilot for the national agricultural census. DANE assessed whether types of variables can be captured through satellite imaging to support information gathered by the census operation.[62] In the same Big Data Survey 2015 mentioned above, both of Mexico's reported projects were on geographical and topographical data derived from satellite imagery.

[61] Secretaría de Turismo 2014.

[62] United Nations Statistical Commission 2014a.

Table 2: Overview of Big Data Projects in Selected LAC NSOs

Type of Big Data	Data Used in Current NSO Activities	Projects	Project Status	Other Organizations Involved
Argentina (INDEC)				
Exhaust data	Web scraping	Building an online consumer price index	Planned	
Brazil (IBGE)				
Digital content	Google Maps	Developing water accounts	Implemented / On-going	
	Call detail records	Tourism monitoring	Planned	
Colombia (DANE)				
Exhaust data	Web scraping	Building an online consumer price index	Planned	
		Price information system (SIPSA)	Implemented / On-going	
Digital content	Call detail records	Monitoring crime activities	Pilot stage	World Bank Data-Pop Alliance TransMilenio
		Socio-economic levels and networks	Pilot Stage	
Sensing data	Satellite	Complementing the national agriculture census		
Ecuador (INEC)				
Exhaust data	Web scraping	Building an online consumer price index		
Digital content	Twitter	Subjective wellbeing		
	Call detail records	Daytime migration		
Guatemala (INE)				
Digital content	Call detail records	Monitoring poverty levels	Pilot stage	World Bank
				Telefónica
Mexico (INEGI)				
Digital content	Twitter	Subjective wellbeing	Completed	InfoTec and Tec Monterrey
		Subjective wellbeing of women	Pilot	Data2x and University of Pennsylvania
		Tourism monitoring	Completed	Ministry of Tourism
		Movements across borders	Planned	

Box 6: Maternal Morbidity and Remote Sensing of Malaria in Brazil

Remote sensing satellite data on vegetation density, soil moisture, population density, and spatial pattern of human infrastructure have long been used to predict levels of malaria risk. Advances in computing now allow more powerful use of these big datasets, including analysis of extreme spatial and temporal heterogeneity and inclusion of greater numbers of explanatory variables. This project seeks to create malaria risk maps for the Amazon Basin, focusing first on urban and peri-urban zones along the Brazil-Guyana border, which are areas with highly variable vector habitats and elevated incidences of illness. At least two vector distribution-mapping studies in this region exist, but to our knowledge there is no high-resolution dynamic mapping of malaria risk. The first phase of the project will use remote sensing data and existing health records, in combination with information about the economic, cultural, and health system, to estimate a spatial regression model that predicts morbidity burden in pregnant women, using DALYs (Disability-Adjusted Life Year) as the principal metric. The second phase will then test the accuracy of this model using data collected in real-time. UN Women and IBGE and the leading institutions piloting this study, drawing support from partner institutions The Vargas Foundation and the Amazon Malaria Initiative.

Today, the Big Data efforts by NSOs in the region, as described above, are isolated projects and remain in the pilot stage. There is nascent, but burgeoning activity with regards to Big Data applications and with organziations taking advantage of others' pilots, approaches, and possibilities. This is not surprising or disheartening. NSOs around the world, even those that have been progressively working on Big Data for years (such as Statistics Dutch), have not yet managed to push Big Data approaches from the pilot status into a functional, sustainable, and wholly implemented operational status. Likewise, in Colombia, Ecuador, Mexico, and throughout the LAC region in general, it is too early to determine how the pilots will be integrated into regular operations. Overall, there are no clear allocations of resources that will galvanize this push, and procuring ample financial and human investment remains challenging.

2.3 Big Data for SDGs across the Wider Ecosystem of Actors

Outside of the statistical system there are various other actors across Latin America that have experimented with using Big Data in an SDG-relevant context. These examples include international Big Data research projects; governments and international agencies; the private sector; and civic technology advocates.

Big Data Research Projects

Most notably, these actors participate in the form of research projects.[63] In Mexico for example, Telefónica and its Telefónica Research Team realized various research projects using CDRs from Mexican customers in order to monitor behavior following natural disasters and disease outbreaks.[64] While Telefónica did use census data for this project, the statistical office INEGI was not actively involved.

In Colombia, the International Center for Tropical Agriculture (CIAT) developed models of agricultural productivity in the context of climate variability. By identifying what type of agricultural practices have historically worked well, in which locations, and during which identifiable spells of weather they worked, the model saved Colombian rice farmers an estimated US $3.6 million dollars in a recent season.[65] In another Big Data research project, scientists from John Hopkins University in the U.S. analyzed health related Tweets as part of Google Trends on flu and dengue outbreaks in Brazil, Mexico, and other countries across the region.

Governments and international agencies

In addition to their Open Government endeavors, Latin American governments have also started to look at Big Data use cases to promote efficiency and improve policies. Social media analysis during election processes, data on traffic flow, and crime prediction are other areas that governments are exploring. The National Roads Institute of Colombia uses GPS data via an electronic tracking device to improve traffic circulation and to serve as input for transport statistics.[66] Whenever a car passes a toll station it is registered automatically. The device also contains all the information about the vehicle, which complements that of the National Single Transit Register. So far, this new method has been tested in 10 toll stations in Colombia and has improved control of traffic flows as well as strengthened transport statistics. In many cities Big Data projects are initiated by multinational companies and implemented as Public Private Partnerships in collaboration with the local government. For example, in the context of Smart City initiatives, such as IBM's engagement in Rio de Janeiro, Brazil, or in Guadalajara, Mexico, the city is analyzing sensor data to improve energy efficiency (among other efficiencies) with Cisco's support.[67]

Other governmental agencies, like Ministries in charge of ICT, communications, or finance, have also evaluated Big Data applications and have started pilots. For example, the Ministry of Finance in Colombia commissioned research using Google Trends to nowcast economic activity (see Box

[63] For an overview of cases of Big Data or data related projects see for example a report from the World Bank "Big Data In Action for Development" and a Working Paper published by ECLAC on "Big Data and Open Data as sustainability tools".

[64] Clark 2013.

[65] CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS) 2014.

[66] United Nations Statistical Commission 2014a.

[67] Pretz 2014.

7). The Ministry of Interior and the United Nations Office on Drugs and Crime used satellite images to measure and monitor coca crops in Colombia through the integrated system of illicit crop monitoring.[68]

As mentioned in section 2.2, donor organizations and international agencies are driving many Big Data efforts in the region. One example is the aforementioned project by the World Bank and Telefónica Research that uses CDR data in Guatemala to estimate poverty. The World Bank conducted other pilots in Nicaragua and Guatemala that test Chen et al.'s approach in using luminosity as a proxy for socio-economic levels (see Annex 5).[69] The World Bank also supports Big Data activities at DANE in Colombia, where this scoping paper has been one element. UN Global Pulse has also played a driving role—as part of its "Rapid Impact and Vulnerability Assessment Fund," UNODC and UN Global Pulse researched how crises may impact crime levels. They focused on four Latin American cities—Buenos Aires, Montevideo, São Paulo, and Rio De Janeiro—using high frequency police-recorded crime data.[70] Data-Pop Alliance, along with Telefónica, and BKF (and funded by The World Bank), has also piloted two initiatives that focus on public safety and crime in Bogotá using CDRs.

Other international institutions, such as United Nations Population Fund (UNFPA) and the United Nations Development Department, with whom most NSOs already have formal agreements, are also becoming increasingly active in this field. The Office for the Coordination of Humanitarian Affairs (UNOCHA) is actively working on improving their Humanitarian Data Exchange (HDX) portal, which will be an interesting source of information and a coordination platform for NSOs. Generally endowed with better resources, those agencies can be important conveners for NSOs in the LAC region (see Table 3). The international alliance CIVICUS focusing on Civil Society Organizations, set up its project Datashift to generate and improve citizen-generated data and fill existing data gaps .[71]

Private sector approaches

Additionally, there are other actors who are working on data-driven innovation around Latin America. As mentioned, multinational companies, such as IBM and Microsoft, invest heavily in Big Data projects throughout the region. Many start pilot projects as Public Private Partnership (in the case of Rio de Janeiro, for example, IBM provides their Smart City technology for free and uses in exchange the data and findings to improve and test their services). Multinational companies, such as IBM and Microsoft, invest heavily in Big Data projects throughout the region. Multinational banks operating in the region have been applying big data applications to identify money laundering and fraud for years.

As in other regions around the world, there are a growing number of start-ups based on data-driven business ideas. Today there are 17 Latin American IT companies whose combined

[68] Pretz 2014.

[69] Chen and Nordhaus 2011.

[70] United Nations Global Pulse 2012.

[71] CIVICUS n.d.

Box 7: Using Google Trends to Nowcast Economic Data in Colombia

The economic indicators used by Colombia's Administrative Department for National Statistics (DANE) to analyze economic activity at the sectorial level have an average time lag of 10 weeks. In order to obtain more real-time estimates of economic activity, the Colombian Ministry of Finance is looking into ways to real-time forecast (i.e. nowcast) activity based on data from Google web searches.

Researchers at the Ministry of Finance analyzed the relative frequency of particular search terms, using *Google Trends*, a tool maintained by Google that keeps track of prevalent search terms over time and provides an index of how common the queries are in each part of the world over a given time period. Based on a methodology for short-term forecasting of economic series developed by Choi and Varian,[a] the researchers used *Google Trends* data to infer economic activity at the sectorial level by choosing certain keywords to act as proxies for consumer behavior,[b] thus providing alternative indicators to traditional statistics in a much timelier manner. These indicators were produced for certain sectors of the economy, such as agriculture, industry, commerce, construction, and transports; other economic sectors such as mining or financial services cannot be assessed with this method.[c]

The resulting sectorial indicators, known by their Spanish acronym ISAAC, were validated against DANE's official indicators of economic activity, and both sets of indicators were made publicly available. The ISAAC data, which pertains to the sectorial level, is aggregated to produce a single leading indicator for economic activity, known as ISAAC+. The project team, led by Luis Fernando Mejía, continues to publish the ISAAC and ISAAC+ on a monthly basis.

A major limitation of such web-based measures is that they risk not being representative in countries where internet penetration is low, as is the case in Colombia (∼ 60%). However, as internet penetration continues to grow, the caveat of non-representativeness becomes less of an issue.

Thus, Colombia's exploration of timelier economic indicators shows promise and has attracted the attention of other countries interested in implementing their own Big Data-based forecasts of economic indicators.[d]

[a] Hyunyoung and Varian 2011.

[b] Mejía et al. 2013.

[c] The World Bank, World Bank Group, and Social Muse 2014.

[d] Mejía et al. 2013.

Table 3: Overview of Big Data Ecosystem in LAC

Actor	Country	Project description
Big Data Research		
Alberto Cavallo and team at MIT	Argentina, Brazil, Chile, Colombia, Uruguay, Venezuela	Useing information from the Internet to generate the Price Index and measure inflation.
Telefónica Municipality of Jalisco	Mexico	Used CDRs from Mexican customers to monitor behavior in the course of natural disasters or after the outbreak of diseases.
International Center for Tropical Agriculture	Colombia	Developed agricultural productivity models for climate variability.
Johns Hopkins University	Brazil, Mexico, and others	Analyzed health related Tweets as part of Google trends on flu and dengue outbreaks.
Monroy-Hernández et al. at MIT	Mexico	Used Social Media and Twitter to examine the drug war.
Coscia and Rios at MIT	Mexico	Used Google data to track Mexican drug trafficking organizations.
Governments		
Government of Guadalajara with support from Cisco	Mexico	Analyzing sensor data as part of a smart city initiative to become more energy efficient, among other goals.
Ministry of Labour	Colombia	Using web-scraping to monitor vacancies in the job market.
Ministry of Finance	Colombia	Using Google Trends to nowcast economic activity in Colombia
Mexico's Office of the President and University of Chicago	Mexico	Detecting maternal death using birth and death, patient discharge records, hospital data, Census data, back to 1990.
The National Roads Institute of Colombia	Colombia	Using GPS data via an electronic tracking device to improve traffic circulation and to serve as input for transport statistics.
World Resources Institute	Colombia	Global Forest Watch using Satellite Data.

(continued)]

(continued)

Actor	Country	Project description
UN Women	Brazil	Remote sensing satellite data on vegetation density, soil moisture, population density, and spatial pattern of human infrastructure have long been used to predict levels of malaria risk.
Ministry of Interior with support form the UN Office on Drugs and Crime	Colombia	Used satellite images to measure and monitor coca crops in Colombia through the integrated system of illicit crop monitoring.
Fundação Getúlio Vargas, Government of Brazil	Brazil	FGV worked with the Brazilian government to use Big Data for economic analysis to make spending more efficient.
International Agencies		
The World Bank and Telefónica Research	Guatemala	Using CDR data in Guatemala to estimate poverty.
The World Bank	Nicaragua, Guatemala	Testing Chen et al.'s approach to using luminosity as a proxy for socio-economic levels.
The World Bank, Data-Pop Alliance	Colombia	Supporting and scoping possibilities for Big Data use for SDGs with NSOs in Latin America.
UN Global Pulse and UN Office on Drugs and Crime	Argentina, Uruguay, Brazil	Researched how crises may impact crime levels using high frequency police-recorded crime data.
Office for the Coordination of Humanitarian Affairs (OCHA)	Latin America	Actively working on improving their Humanitarian Data Exchange (HDX) portal.
Civic Technology Movements		
Open Intelligence	Mexico	Helped the Mexican Ministry of Interior to understand neighborhood crime rates based on various data sets.
SocialTIC	Mexico	Supports the government in implementing their Open Data strategy and organize community events to make use of government data.
Unidos pela Segurança (UPSEG) developed by Stal IT	Brazil	Uses crowd-sourcing to allows citizens to report criminal incidents and contribute to public safety.

(continued)

(continued)

Actor	Country	Project description
Multinational Banks	Latin America	Applying big data analysis to identify money laundering and fraud.
Private and Start-up Actors		
IBM		
Microsoft		
Random Monkey (formerly Aentropico)	Colombia	Analyzes big data.
Cignifi	Brazil	Analyzing patterns in mobile device use to predict someone's lifestyle and his/her corresponding credit risk profile.
BogoHack	Colombia	Organize science hacks and hackathons.

value is more than US $250 million, and many of them have developed just within the past four years, stemming mainly from Brazil and Argentina.[72] According to a study by Frost & Sullivan, Brazil, Mexico, and Colombia have already invested in Big Data analytics resulting in revenue of US$603.7 million in 2014 alone;[73] the data analytics company Aentropico (now Random Monkey) is one of such pioneering businesses.[74] The Brazilian start-up Cignifi analyzes patterns in the usages of mobile devices to predict someone's lifestyle and his corresponding credit risk profile. It focuses on the 100 million middle-class citizens who have limited access to financial products such as mortgages or loans due to a lack of traditional credit data.[75] In Mexico, the start-up Open Intelligence has developed a platform that analyzes government data and supports governmental bodies to use their own data for evidence-based decision taking (See Annex 7).

Civic Technology Movement and Critical Voices

Many Latin American countries, such as Chile, Argentina, and Brazil, have seen strong Open Source movements and an overall interest in social "hacking" spurred by Open Government approaches. We find several civil society organizations at the intersection of Civil Tech and data that organize hackathons and science hacks. In Mexico, the NGO SocialTIC supports the government in implementing their Open Data strategy and organizes community events to make use of government data. In Colombia, hackers from BogoHack organize science hacks

[72] Téllez 2015.

[73] Campos 2015.

[74] Campos 2015.

[75] Datafloq 2015.

and hackathons and GeoCensus focus on the application of geodata. In addition, data crowdsourcing projects throughout the region support citizen action; platforms like CIC by Citivox in Mexico or Unidos pela Segurança (UPSEG) developed by Stal IT in Brazil, allow citizens to report criminal incidents and to contribute to public safety.[76]

In addition, to the Civic Technology movement there is a growing number of Civil Society Organizations and research institutions that are actively involved in a critical discourse around data in the hands of governments and private companies. Most of them are part of trans-continental and international networks advocating for Human Rights in a digital age, including the Right to Privacy. Among those organizations are Fundación Karisma, Colombia, la Red en Defensa de los Derechos Digitales, Mexico, Derechos Digitales, Chile or Instituto de Tecnologia & Sociedade do Rio, Brazil. They will be important voices and advocates for the interests' of citizens and consumers in an emerging Big Data ecosystem.

Table 4: Civil Society Organizations Working on Digital Rights

Organization	Country
Fundactión Karisma	Colombia
R3D, Red en Defensa de los Derechos Digitales	Mexico
Derechos Digitales	Chile
Universidad de Palermo, Centro de Estudios en Libertad de Expresión y Accesso a la Informaci ón	Argentina
Asociación por los Derechos Civiles	Argentina
FGV Direito Rio	Brazil
Colnodo	Colombia

2.4 International Attempts to Use Big Data for Official Statistics and Development

UNECE and (more recently) UNSD have specifically driven discussions on potential use cases of Big Data for NSOs, as well discussions on the implications of Big Data, in general. A High Level Group for the Modernisation of Statistical Production and Services was set up in 2010 to oversee and coordinate international work relating to standard-based statistical modernization; and in 2014 UNSD created a global working group (GWG) on Big Data for Official Statistics, whose mandate is based on strategic considerations with particular links to the Post-2015 Development Agenda, the "Data Revolution" initiative, and the Fundamental Principles of Official Statistics.[77] These give valuable insights into potential areas where Big Data could be used to measure the

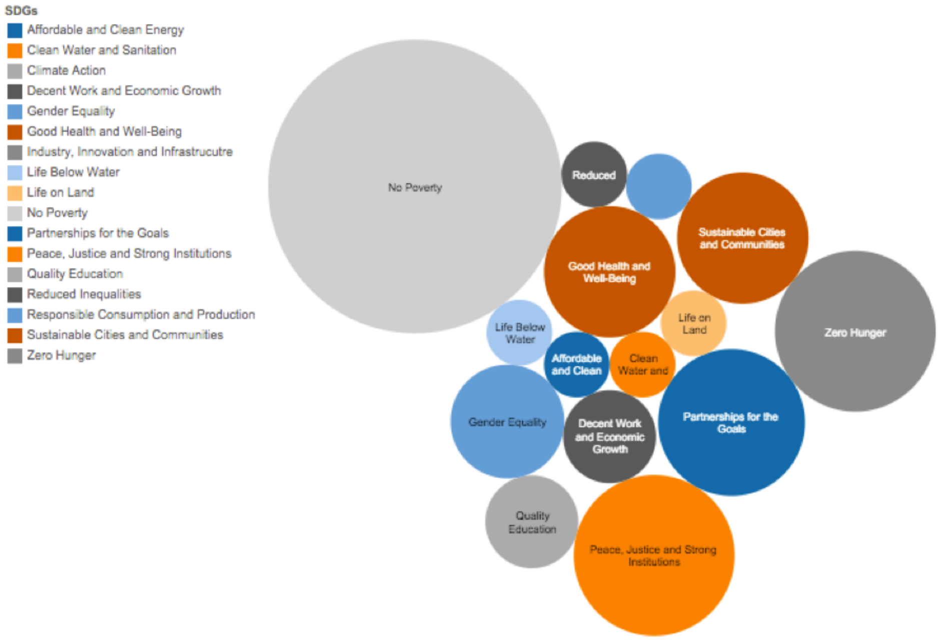
[76] Diniz, Girard, and Perini 2013.

[77] United Nations Statistical Commission 2014b.

SDGs.[78]

Below is a figure using 2015 data from the World Bank's survey on Big Data initiatives for SDGs that shows all the SDG targets which organizations are targeting from around the world (from the LAC region, only INEGI and IBGE responded). [79]

Figure 5: Comparison of Self-Reported SDG Targets



Data from 2015 World Bank survey on Big Data projects for the SDGs.

Across the globe NSOs have started to work with big data sources and to slowly engage with the wider Big Data ecosystem. The Statistical Institute of the Netherlands conducted several pilots, including an analysis of traffic, CDR, and social media data to predict subjective well-being. Italy and the Netherlands have both used mobile phone data to monitor mobility

[78] The High-Level Group for the Modernisation of Statistical Production and Services is sponsoring a series of international collaboration projects to better understand how to harness the power of "Big Data" and other new data sources, to support the production of official statistics. This work supports the concept of a "Data Revolution for Sustainable Development" and the development and monitoring of new sustainable development goals. These projects are open to all national and international statistical organizations that want to contribute.

[79] In responses from the World Bank survey on targeting the SDGs, INEGI in Mexico targeted SDG 10.7 (Facilitate orderly, safe, regular and responsible migration and mobility of people, including through implementation of planned and well-managed migration policies) and 17.9 (By 2030, build on existing initiatives to develop measurements of progress on sustainable development that complement GDP, and support statistical capacity); and IBGE in Brazil reported targeting 1.a (Ensure significant mobilization of resources from a variety of sources, including through enhanced development cooperation, in order to provide adequate and predictable means for developing countries, in particular least developed countries, to implement programs and policies to end poverty in all its dimensions).

statistics. China and the UK (ONS) have conducted research projects on the use of Big Data for pricing, and on analyzing smart meter data for identifying household structures.[80] Others have used mobile data for daytime population, mobility, and tourism statistics, among others. However, as remarked by Statistics Netherlands, the official statistics community is only scratching the surface when it comes to exploring Big Data[81] and many claim that NSOs will have to undergo some radical paradigm shift in statistical methodology, in order to let Big Data gain ground in official statistics.[82]

3 Challenges and Requirements for NSOs Engaging Big Data for SDGs

As stated earlier, a number of operational challenges, such as access to administrative records, already hamper Latin American NSOs in their ongoing statistical activities; many of these same challenges also limit their potential to engage with Big Data. This section addresses the most significant challenges NSOs face in engaging with Big Data and provides recommendations on approaches and next steps NSOs can take to address these issues.

Through literature review, interviews, case studies and SWOT analysis (Annex 10), we identify five major challenges for Latin American NSOs engaging with Big Data: institutional barriers to innovation and change management; constraints on data access and completeness; technical challenges; human capacity gaps; methodological challenges; and legal and political risks, which are discussed in turn.

3.1 Institutional Barriers to Innovation and Change Management

Using Big Data is a significant undertaking for an NSO. It is likely to involve a culture change, requiring both increased interactions with external examples and actors as well as internal willingness toward innovation and transformation. Latin American NSOs face institutional barriers to innovation and change management largely for due to a lack of internal digital culture and a skeptical outlook on new data sources.

Lack of internal digital culture and linguistic skills

There are certainly encouraging examples for NSOs' willingness to transform towards innovation and openness. DANE's innovation process (a part of DANE Moderno—see Box 4) presents a unique example for providing space for innovation in the region. Additionally, NSOs have more

[80] Instituto Nacional de Estadística y Geografía (INEGI) de México 2015.

[81] Daas and Loo 2013.

[82] Scannapieco et al. 2013.

directly engaged with citizens through social media and provided infographics as visualizations of their latest reports. However, despite these new efforts, it will take time to see the results of such initiatives in terms of a genuine cultural change. Overall, NSOs remain conservative towards innovation and change; standards and quality—not innovation and experimentation—define good statistics. A hurdle is also the fact that many resources and dialogues are in English, which some staff do not master well, understandably.

This apprehensive culture is reflected in rather analogue internal practices. In many NSOs in the region, staff are still not allowed to access the Internet from their offices, which can only be partly explained by confidentiality restrictions. If the Internet and new technologies are barred from everyday work life, a cultural shift with NSOs towards new, Internet-derived sources remains significantly difficult.

In addition there is general skepticism towards new data driven approaches by the staff of NSOs in the region as they see new technologies and Big Data as potential threats to their jobs. With a long tradition in household surveys, NSOs in Latin America are employers to several thousands of people, and are hesitant to embrace new sources of data with which they are not familiar. Often this is also due to a general lack of understanding and misconception of Big Data, which is mistaken with more general IT projects, the use of social media (as in social media monitoring), the building of data warehouses, and recent activities in the field of open data. This issue needs to be taken into account and employers on all levels need to be informed about actualities and implications when engaging in new projects involving Big Data.

If digital processes can be adopted internally, it is more likely that an organization and its employees will recognize their value. Here it needs sufficient backing from the top and high-level commitment towards these developments. At the same time hands-on approaches, trial and error pilots on Big Data and exchange with colleagues from other NSOs across the region can stimulate acceptance, understanding and interest in Big Data approaches among employees and help to illustrate use cases. For example, in the case of an internal hands-on experiment in Mexico, staff at INEGI were able to receive a first glimpse of the actual value of Big Data applications for their operations and, at the same time, acquire a general understanding of how this might change their work in the future for the better. This will be crucial to also ensure that pilots and projects will be driven by local priorities, suitability and embedded in regional debates, including about potential risks and challenges.

Recommendations

- Promote digital working culture
- Actively inform staff about Big Data application and implications
- Allow pilots and hands-on use of Big Data sources to test potential use cases
- Develop resources in Spanish and Portuguese, develop English skills, and encourage multi-language content and exchanges

3.2 Constraints to data access and completeness

The private sector is deeply engaged throughout the data value chain and much of the data revolution largely hinges on the inclusion of companies.[83] However, there remains limited cooperation across the LAC region between NSOs and the private sector, particularly the telecommunication industry. Currently, the exchange with the private sector has been mostly determined by general agreements on the exchange of data in the context of traditional statistics (i.e. NSOs requesting a company's data for completing registers). However, public-private partnerships and other forms of collaboration for the exchange of knowledge and skills are fairly new forms of engagement for NSOs.

Accessing private sector data therefore forms the hardest part for proceeding with Big Data in Latin America; this is also reflective across other regions as demonstrated in a recent UNECE survey.[84] As discussed earlier, some forms of Big Data such as social media data (e.g. Twitter data) are partially available and could be a promising source to monitor and improve socio-economic data for measuring SDGs. However, other forms of Big Data (e.g. CDRs) remain strictly kept on company servers. While several Latin American NSOs have expressed interest in working with CDR data, the limited possibility to access private data sources has slowed down these initiatives. Telefónica has been the only mobile operator openly using its data for research purposes in Latin America. Given the fact that América Móvil (via its subsidiaries) has substantial mobile penetration throughout the region, it would be crucial to understand potential incentives for collaboration and opening América Móvil CDR data.

Furthermore, even ongoing research pilots using CDR data have been widely steered by mobile operators rather than NSOs. In Mexico, Telefónica has done research and used data from INEGI, and yet the institution has not been involved in the research. A current research project based

[83] Ballivian and Hoffman 2015.

[84] United Nations Economic Commission for Europe (UNECE) 2013.

on Telefónica data in Guatemala was established as a result of the convening power of the World Bank as a third party. In Brazil, IBGE has unsuccessfully tried to access CDR data for pilot purposes and have now asked the national telecommunications agency (ANATEL) for assistance. It remains to be seen whether they will be successful. Occasional access to data can be a first step towards engaging and experimenting with data, but will lead to little sustainability. This is also true in the context of social media data when access to data via API is discontinued.

Past research on CDR-data in other regions could be done because it was either done in-house (with Telefónica Research), framed in specific agreements between research institutions and the operator (e.g. in the Netherlands and in Italy, NSOs had agreements with Telekom)[85] or under special arrangements as part of "data philanthropy" approaches,[86] which also involved the setup of a formal agreement (Orange D4D Challenge).[87]

As of now there is simply no coherent and comprehensive set of regulations or guidelines that govern the access to CDR or other data from the private sector (as mentioned it is to expect that for example Whatsapp data hold by Facebook will become more important). It lacks openly available, easy-to-use and legally compliant resources for establishing these partnerships, which need to be cross-industry and cross-jurisdictional.[88]

To fully benefit from Big Data, private corporations, NSOs, and governments need a "New Deal" on data.[89] The SDGs illustrate that new and international policies that change current approaches for accessing and using data are needed. The monitoring will require stable and sustainable access to data on a global scale. It is unlikely that single NSOs or governments can enforce such shift; instead, they will require global agreements, i.e. supported by the UN or the World Economic Forum.

This new deal will hopefully be grounded in broader public debate on data ownership. Approaches like data philanthropy, for example, imply too strongly that the data belong to mobile operators and not to the individual data emitters of data. Various academics like MIT or civil right organizations like the Open Knowledge Foundation challenge this approach.[90] So while it will be important to build up solid partnerships with the private sector, NSOs should not become solicitant of private companies or their project ideas. The interest of a private company, driven by economic incentives, will seldomly be congruent to the interest of an NSO that has a public agenda. So far, this debate is still weak in Latin America. Hopefully the region can benefit from a global discourse, as Latin American civil society organizations increasingly

[85] United Nations Statistical Commission 2014a.

[86] Pawelke and Tatevossian 2013.

[87] Orange 2014.

[88] Ballivian and Hoffman 2015.

[89] Pentland 2009.

[90] Pentland 2009.

become engaged in the discussion.

Recommendations

- Engage with Private sector
- Evaluate current models for corporate data sharing
- Set up agreements for public-private Partnerships

3.3 Technical Challenges

Big Data poses a set of technical challenges and obstacles particularly in terms of quality control of statistical processes. The volume of data requires an expansion of processing techniques matching modern hardware infrastructure and re-engineering storage systems. Learning algorithms require appropriate computing capacities for the variety of the data calls that allow the combination of different data types collected at different levels, sometimes with temporal or geographical structure.[91] Unstructured data (such as satellite data and social media data) requires specific analytical capabilities in order to manually train algorithms to classify this content automatically. Structured data such as CDR data can be easier to process, but often need to be validated with other sources such as satellite data or household services.

These challenges are in fact similar to the challenges involving use of administrative data. Not surprisingly, NSOs across the Latin American region still struggle with the technical transition to the increasing use of administrative data composed of structured and unstructured data, which requires new standards and formats. Many NSOs are currently in the process of building data warehouses as central repositories of databases that focus and integrate surveys, censuses and administrative records. These ongoing activities will help improve the activities of capturing, cleaning, processing, analysis and visualization of data by using tools to have automatic controls in processing, with standardized variables and databases of the same theme.

In addition, the strong Latin American Open Data movement favors Big Data efforts as it encourages standardization processes data formats. The transition towards open and exchangeable data formats, such as the OECD standards for micro- and metadata (csv and sdmx), eases some big data applications. ECLAC, for instance, already supports the harmonization of software, technologies and tools, including the methodological harmonization for sharing data among the region,[92] which already smoothens the SDG measuring process and improves data quality.

[91] Kreuter and Peng 2014.

[92] Economic Commission for Latin America and the Caribbean (ECLAC/CEPAL) 2013.

Specific IT tools and techniques will need to be adopted to embrace Big Data. The huge size of data sets requires using distributed file systems to overcome physical limitations. Platforms for managing complex storage systems, such as Hadoop, are therefore required. These major information technology components are frequently used in the process of collecting, storing, and analyzing Big Data include (see Annex 9).

Recommendations

- Use freely available services
- Share tools and software among NSOs
- Promote and benefit from Open Data efforts

3.4 Human Capacity Gaps

In developing regions, a lack of trained statisticians still generally poses a major challenge for most NSOs (INEGI again constituting an exception). In many Latin American countries, in the fight for young statisticians, NSOs compete directly with international organizations like UN chapters and cannot keep pace with salary. This issue becomes even more striking with the rise of Big Data demanding very specific skills. Data and computer scientists, for example, who are able to manipulate complex datasets, and data engineers, who design the IT architecture for collecting and processing data, are scarce.

In Latin America, so far only a few universities offer training in Data Science, and NSOs compete with start-ups and Internet companies in the fight for data experts. However, in the already more prosperous countries this seems to be changing, and the number of master's programs is on the rise, for example, at the University of Los Andes in Colombia or at the Instituto Tecnológico Autónomo de México in Mexico City. In Brazil, there are actually a high number of graduates in the area of computer science.[93]

In-house training programs could offer a solution: Some NSOs, like in Peru, Colombia and Mexico, offer their own training programs to educate their own staff and others on new methods. CANDANE, the training section of DANE, was founded some years ago and currently trains around 1500 students in basic statistics, questionnaire design, and the use of tools like Stata and SAS. This is done via direct training and via online e-learning tools, which are also offered to students from other countries and continents. There are some initial ideas to also offer data analyzes in collaboration with university partners. The promotion of e-learning tools and webinars are considered helpful instruments; unfortunately most training programs

[93] Digiampietri et al. 2014.

currently offered online are offered in English, producing barriers to learning The set up of Spanish or Portuguese courses could be an easy way to not only promote expert knowledge on Big Data but also to inform the wider community in the statistics office.

Another approach is to give scholarships to staff for specific training classes, as INEGI does. Outsourcing, for example to IT companies and university institutions, as well as insourcing, for example by hiring summer interns or employing people on a project-by-project basis, have been further successful approaches in Mexico. Yet bureaucracy often makes it difficult to hire someone short-term or ad hoc in other countries.

Universities and academic institutions are also relevant partners. As mentioned, an increasing number of universities are entering the field of master's programs in Computer Science. However, until today, collaboration with universities commonly has been based on formal agreements for data exchange and not on the strategic exchange of knowledge. Although NSOs usually provide a processing room for researchers, such spaces are hardly used because of excessive bureaucracy or old technical infrastructure.

But it would be problematic to focus only on technical skills (which is currently happening in most NSOs), as other skills are just as important. Because of the interdisciplinary nature of Big Data, effective use of Big Data requires multi-disciplinary teams, ranging from:[94]

- *Domain Expert.* A user, analyst, or leader with deep subject matter expertise related to the data, their appropriate use, and their limitations.
- *Researcher.* Team member with experience applying formal research methods, including survey methodology and statistics.
- *Computer Scientist.* Technically skilled team member with education in computer programming and data processing technologies.
- *System Administrator.* Team member responsible for defining and maintaining a computation infrastructure that enables large scale computation.[95]

At INEGI, partnerships with universities have long been established and therefore support current initiatives for the use of Big Data. In a current pilot, INEGI has started to analyze Tweets to understand subjective well-being in Mexico. This project has been set up with an interdisciplinary team of researchers, ranging from computer scientists to linguists.

Informing internal staff will be crucial to build up capacity on Big Data, in particular, where there is general skepticism against new technologies. This includes explaining the interrelation and differences with other approach such as Open Data. Intranet and internal magazines could be other tools to raise awareness.

As stated above, the gaps between rural and urban settings still from one of the major political but also statistical challenges in many Latin American countries. Therefore, it will be crucial to

[94] American Association for Public Opinion Research (AAPOR) 2015.

[95] American Association for Public Opinion Research (AAPOR) 2015.

involve local entities, for example at the municipal level, in the debate about Big Data. Municipality structures are still very weak, and many of them do not use the data or collect qualitative datasets. While data revolutions offer broader opportunities for bigger cities and controlling bodies (like the NSOs), it remains unclear how local actors can actually influence and benefit from these developments.

Recommendations

- Partner with local universities
- Use alternative forms of training, such as web-training
- Promote data literacy among staff

3.5 Methodological Challenges

Statistical quality formulates one of the key principles of NSOs (see also Principles 3 of the UN Fundamental Principles of Official Statistics). The fabric and distribution of Big Data, however, demands different processes than traditional statistical sources to meet these quality standards. Big Data itself poses challenges in terms of representativeness.

As a data source, Big Data is generally not designed to answer specific scientific research questions, but is used for other purposes than they are collected for: inference (solid sampling process) and measurement (covering all relevant variables).[96] These discussions are not entirely new to the statistical community, as similar questions have been raised with the use of administrative data. This is slowly enforcing a new paradigm shift, in which the number of design-based approaches originally used within Official Statistics is decreasing. But for administrative records, NSOs can at least advocate or influence ministries and agencies generating the data to design the registers accordingly. This will be difficult for web or social media data. Model-based approaches are difficult to apply to Big Data analysis. Approaches that proceed by exploratory analysis, like those based on data mining and machine learning, could be applied more appropriately.[97]

In addition, the process of analyzing data introduces risks "for noise accumulation, spurious correlations, and incidental endogeneity which may be compounded by sampling and non-sampling errors. Related to the former, data may be filtered, sampled or otherwise reduced to form more manageable or representative data sets. These processes may involve further transformations of the data. Errors include sampling errors, selectivity errors (or lack of representativeness), and modeling errors."[98] For that reason—while the collection is cheap—Big

[96] Kreuter and Peng 2014.

[97] Eurostat 2014.

[98] American Association for Public Opinion Research (AAPOR) 2015.

Data can be expensive to clean and process, requiring a more human capital for structuring, linking and managing the new types of data.[99]

The lack of representativeness of Big Data today forms one of the major challenges. Even Big Data streams with huge N are most often not representative of entire populations. The populations covered by Big Data sources are not typically the target populations of official statistics and are often not explicitly defined. Also it is not always feasible to assess the relationships between the covered population and the target population, on the one side, and to estimate the bias, on the other.[100] This is particularly striking in the context of SDG and poverty measurement in particular, as biases in the data might neglect particular groups that need to benefit from the post-2015 agenda, like indigenous groups, women, groups with low-income levels etc.[101]

For the LAC region this bias presumably rests in the gap between rural and urban regions, as is reflected in mobile phone penetration, and which often also reflects socio-economic biases between different groups and minorities. Hence, the likelihood that those that have not been covered sufficiently by traditional data will not reflected adequately in big data is high or even higher. For this reason, it needs to be assured that CDRs can actually be used to monitor the targeted population, and to judge if CDR data is an accurate analysis tool for urban areas only. The same is true for social media data, which is widely distributed in Brazil, Chile, and Mexico, but does not have the same level of availability in other Latin American countries.

There can be other reasons for biases. In the case of Colombia, many people do not use their own phones to make calls, but use the phones of so-called "minuteros"—sold by street vendors of fruits or other goods and at the same time offer "minutes" on different phones. This is cheaper for those who need to make calls to different providers, which is usually still very expensive. Considering the CDR data that is produced by the phones of these "minuteros," it is easy to imagine that they will have an interesting, yet confusing, output for researchers.

For traditional data there are several frameworks to mitigate errors in the survey process. For Big Data, the most likely solution will be a "combination off traditionally designed data and Big Data. However, such solutions of data linkage and information integration are themselves threatened by concerns about privacy and confidentiality."[102]

Researchers and statisticians in Europe are already investing in techniques to avoid and detect data biases. A lot of investments and work will need to go into developing such methods if Big Data is to be widely and used for monitoring purposes. Generally Big Data can be just as good as the data against which it is controlled.[103] Certainly, solid ground-truth data like Census and survey data, or satellite data, are needed to detect biases. Those are not always available on all levels in the Latin American region. The region might also need distinct approaches in adjusting current methodologies, specifically to address the issues of gaps between rural and urban settings. A first step could be to improve ICT statistics on a more granular level, which could be done

[99] American Association for Public Opinion Research (AAPOR) 2015.

[100] Eurostat 2014.

[101] boyd and Crawford 2012.

[102] Kreuter and Peng 2014.

[103] Smith, Mashhadi, and Capra 2013.

by analyzing mobile phone distribution in rural areas in order to get a better understanding of potential biases.

Similar to any statistical processes, standards and guidelines on an international level will be needed, both to guarantee the quality of the data and to allow comparability (see above). The LAC region could play an active role in fostering these standards on an international level: via the various UN groups, the World Bank and other international agencies, and by putting the their region-specific challenges (bias between rural and urban areas) on the agenda.

Recommendations

- Continue further investments in Big Data research
- Consult ongoing global conversations on Big Data and measurement
- Initiate the development of new standards and guidelines in the region

3.6 Ethical and Political Risks

While NSOs naturally have more expertise in dealing with confidential data than many other institutions, potential privacy and data protection risks are indeed much higher in the context of Big Data and digital data in general. The challenges attached to Big Data range from a lack of ownership of the data, purpose limitation (to the definition of official statistics) and the limits of data anonymization in the context of digital processing of data sets.

Since the data is not generated by NSOs and in most cases not even generated for statistical purposes, there is a lack of clear legal frameworks. Most consumers of digital services (such as smartphone applications) and hence emitters of Big Data have little or no idea that their data may be re-used for other purposes, such as statistical products.[104]

In the digital age the anonymization of digital data sets is limited. In contrast to the most common tool used in statistical processes, removing personally identifiable information (PII) would hence not be sufficient to protect against re-identification.[105] Crossing certain data sets with similar data easily allow re-identification of individuals, and only a small number of data points (for example position, date, and time) are needed to re-identify an individual in the dataset.[106] In particular, location data, such as tourism or migrations statistics, and which could be highly valuable for statistical products, poses enormous risks to anonymization. Generally, aggregated data, such as Antenna-to-Antenna traffic (as has been done by Smith et al. in their research on poverty

[104] American Association for Public Opinion Research (AAPOR) 2015.

[105] Montjoye et al. 2013.

[106] Montjoye et al. 2013.

estimation in Senegal) would not intrude on privacy. But as researchers from EuroStat point out, the aggregation of initial data before it is processes, seriously limit the available options in relation to methodology and the data's potential for statistics.[107] While there have been technical and methodological attempts to solve the issue on a technical level, for example by adding noise to the data to make the re-identification more difficult, many technical scholars argue that thorough de-identification can never be guaranteed.[108] Among Latin American NSOs there is only little awareness about the limits of traditional methods to anonymize data in the context of digital data.

As Big Data is not produced by NSOs in-house but demands new partnerships with the private sector, this also changes the legal baselines of the use of personal data. Today many countries in Latin America do have Omnibus Data Protection laws closely modeled after EU laws: Argentina, Uruguay, Mexico, Peru and Costa Rica are at the forefront of privacy; Brazil, Colombia and Chile recently followed suit.[109] These laws reveal similar concepts, such as the EU directive including special treatment of sensitive data, notice and consent, obligation to keep the data secure, restriction for cross-border transferring, creation of data protection authorities, registries of data bases, and data subjects' right to access and control their data.[110] The concept of Habeas Data forms the baseline for most of these laws. The Habeas Data right "[we command] you the data," builds on principles of the German "Right to Informational Self-Determination" and the Council of Europe's 108th Convention on Data Protection of 1981, protects the personal information by allowing that person to request the rectification, update or even the destruction of the personal data held in (an automated) database, and implies that there must be transparency on the gathering and processing of such data.[111] Generally, there are other laws that also affect the use and control of data sets, such as Colombia's law on personal data (Law 1581 from 2012).[112] This legislation states that personal data can be classified as either personal or semi-personal data (under Habeas Data or Law 1266 from 2008)[113] and sensitive or non-sensitive data (under Law 1581 from 2012).[114] However, the definition extends not only to data considered as personal data or sensitive information, but also to other data which also deserve to be preserved, controlled and considered equally worthy of preservation, control and disclosure.[115]

However, in Latin America law enforcement is still very limited.[116] Even where law enforcement is strong, current privacy frameworks—across the world—poorly address the illustrated privacy challenges caused by Big Data. Since the data is usually passively collected, other policy tools that are based on informed consent and purpose remain problematic because the potential use of the data might not be yet defined when the data was collected. This will be even more difficult when

[107] Eurostat 2014.

[108] Ohm 2010.

[109] Martinez-Herrera 2011.

[110] Martinez-Herrera 2011.

[111] Guadamuz 2016.

[112] Congreso de la República de Colombia 2012.

[113] Congreso de la República de Colombia 2008.

[114] Newman Pont 2015.

[115] Urioste Braga 2009.

[116] Martinez-Herrera 2011.

the subjects are those living in poverty or are highly vulnerable[117] and where legal frameworks might not be entrenched at all. For that reason it will be essential to develop international legal frameworks for using new data sources in an appropriate, and value-driven manner on the one hand; and on the other hand to remind NSOs of their ethical standards and responsibility towards the public good.

After all, also in a Big Data context the Fundamental Principles of Official Statistics continue to provide the ethical guidelines for NSOs' activities. This would for example also apply to questions around the minimization of biases in the data and NSOs responsibility to inform the public about those impediments.

As mentioned before, NSOs in many countries are seen as trusted actors in handling and controlling data, which could support their potential roles as relevant third parties in the Big Data ecosystem. Clearly it will be essential to ensure confidentiality for safeguarding trust, both for new activities but also for on-going work such as surveys. Governed by legislation that put the protection of data at the heart of the statistical process, NSOs would be good candidates to move the Big Data discussion towards good practices for societal well-being.

But as indicated earlier, the quality of legal frameworks of NSOs across the LAC region differ. Many NSOs still do not yet follow the best practices recommended by the UN such as independence. This impairs their potential role in the Big Data ecosystem as trusted third parties, as it might be more difficult to convince societies that they could fulfill this role. In times of on-going mass surveillance, it will also be important to illustrate that NSOs are not new governmental tools for effective surveillance. Several events in Latin America have incited distrust among citizens and civil society activists. During the protests against the World Cup in 2014, the police and intelligence services in Brazil surveilled protesters. In Mexico, the last amendment of the Telecommunications Act included explicit policies regarding the geo-location data of cell phones without requiring a court order.[118] In Colombia, different actors have been put under surveillance during peace negotiations.[119]

NSOs in the region need to be cognizant of this challenge and promote a value-driven Big Data approach. The impacts of Big Data and the limits of anonymization also need to be integrated in the NSOs' Code of Ethics and Good Practices. Additionally they need to actively inform the public of the risks and benefits of Big Data. The involvement of civil society groups, human rights advocates, journalists, and privacy activists will be essential to the development of valid legal frameworks. "There is a need for more thought and discussion on the shared risks, incentives and impacts for establishing multi-stakeholder data sharing agreements."[120] Transparency about Big Data activities and partnerships will certainly be key to promote trust. NSOs could even become key players in providing the public with data and promoting transparency about other Big Data activities conducted by government agencies, as suggested in a recent study on the risks

[117] Ballivian and Hoffman 2015.

[118] Ruiz 2014.

[119] Barbosa 2014.

[120] Ballivian and Hoffman 2015.

of Big Data usage in the labor market in Chile by Derechos Digitales.[121]

Despite the above-mentioned incidents, some Latin American countries have actually been pioneers in multi-stakeholder approaches. Brazil, for example, has been at the forefront of promoting human rights-based approaches to the Internet and protecting freedom online—most notably by pushing for a new international declaration on the right to privacy in the digital age. The Brazilian Marco Civil da Internet has been globally recognized as an example of a best practices process for multi-stakeholder engagements, including its involvement of the private sector.

- Development of privacy assessment tools
- Development of ethical frameworks surrounding Big Data
- Consider multi-stakeholder approaches
- Norms and laws about *use* of data

4 Towards a Regional Multi-Partner Roadmap for Leveraging Big Data for Official Statistics and the SDGs

4.1 Five Regional Trends Promoting Big Data Use in Latin America

NSOs remain a pivotal actor in the ongoing evolution of official statistics and the achievement of the SDGs during the data revolution, both within their own mandated activities as well as in the formation and development of the regional ecosystem of actors using traditional and new data sources. To fulfill this role, NSOs need to actively engage the Big Data ecosystem to ensure that the yet-to-be-defined path of Big Data leads towards societal progress. The measurement of the SDGs will be an important task for the next fifteen years, and there is certainly evidence that Big Data could help NSOs fulfill this responsibility. Additionally, the SDGs are to serve, for the first time, as global indicators that involve all countries. "The world we want," the subtitle of the first report of the UNDG Millennium Development Goals Task Force, goes beyond better numbers and measurements. For that reason, it would be a big mistake to believe that Big Data would only be about new data sources— it will and does have wider implications for the texture of societies. To avoid a second digital divide, developing regions need to have a say in this discussion, and NSOs are the most opportune and some of the most capable actors to coordinate this process.

[121] Velasco and Viollier 2016.

As we described in the previous section, major challenges and barriers persist for NSOs to leverage Big Data:

1. **Institutional barriers to innovation and change management**, including the lack of an internal digital culture, skeptical institutional outlook on new data sources, and lack of coordination among stakeholders;
2. **Constraints to data access and completeness**, particularly in access and continued use of private sector data, lack of public private partnerships and limited ownership rights involving people and their relationships with data;
3. **Technical challenges**, including infrastructure for capturing, cleaning, processing, analyzing and visualizing both structured and unstructured data as well as adoption of specific IT tools and techniques;
4. **Human capacity gaps**, including talent discovery, data literacy, limited data science training programs, and limited involvement of universities and other academic institutions;
5. **Methodological challenges**, including challenges in data representativeness, biases, and the lack of standards and guidelines;
6. **Ethical and political risks**, including risks to privacy and weak legal frameworks.

Despite these challenges, we derive the following major regional trends that, in addition to the SDGs, facilitate further Big Data use and experimentation across the Latin American data ecosystem:

Latin American Experience of the Open Data Movement

The Open Government and Open Data movements have generated significant political capital in Latin America around data for public good, particularly in Peru, Mexico, Colombia and Brazil. The region retains the highest proportion of country participants—fifteen member countries—relative to other regions in the Open Government Partnership (OGP), a global alliance to promote Open Government. The majority of these member countries have initiated national action plans on citizen engagement, transparency and government accountability. Advocates in the Open Data and transparency movements—journalists, academics, infomediaries and communities of civic hackers—have pushed for the transformation of existing government data into machine-readable, accessible formats for research, analysis and advocacy. For example, Mexico's open data legislation has official Open Data laws that have already led to valuable developments such as the transition towards Open and exchangeable Data standards (SDMX); this transition favors both Big Data activities and the efficient measurement for the SDGs. CEPAL's analysis of the Latin American data ecosystem points to the synergies created by both the Open Data and Big Data movements.[122]

[122] Economic Commission for Latin America and the Caribbean (ECLAC/CEPAL) 2014.

The Emergence of Public-Private Partnerships on Big Data

As noted earlier, the emergence of public-private partnerships on Big Data is a fairly recent development in Latin America. Private sector companies working in partnership at some level with non-private sector entities on data-related activities include Telefónica, IBM, Microsoft, multinational banks, Aentropico (formerly Aentropico), Cignifi, and Open Intelligence. While limited cooperation across the region often inhibits formal partnerships between NSOs and the private sector in some industries, public entities have been able to access data that private sector companies have shared through prizes and challenges, APIs, and intelligence products. For example, geolocated Twitter data for INEGI's work on subjective well-being was derived (in partnership with academic institutions) through Twitter's public API. Similarly, DANE and researchers at the Ministry of Finance used data from one of Google's intelligence products (Google Trends) to infer economic activity across sectors. While these forms of data sharing represent limited forms of partnership (relative to more formal public-private partnerships), they do represent an emerging phenomenon of corporations sharing data through different shades and models of openness.

The Presence of Strong Region-wide Committees, Institutions and Working Groups

Existing structures within the LAC regions allow NSOs to foster Big Data as a source for advancing SDGs. For example, ECLAC supports the harmonization of software, technologies and tools, including the methodological harmonization for sharing data among the region,[123] which already smoothens the SDG measuring process and improves data quality. The Statistical Conference of the Americas of ECLAC promotes the development and interoperability of national statistics in the region for international comparative analysis, as well as cooperation among NSOs at bilateral, regional and international levels.[124] ECLAC helps facilitate regional working groups for NSOs and other statistical actors, particularly on cross-cutting thematic issues such as gender statistics, migration and trends in remittences, MDG progress, and environmental statistics.[125]

The Development of Adaptable Best Practices

Most LAC countries face similar challenges, and at the same time could benefit from best practice examples from the region. Right now, NSOs in Colombia, Mexico, Ecuador, and Brazil are actually dealing with the same issues and want to start similar pilots; some of them have already built the required software and tools while others have investigated in the methodology. Mexico's INEGI has hosted fifteen international meetings to date on best practices for gender statistics across the region—both in the "production and in the use of data for the preparation, implementation, monitoring and assessment of public policies, as well as

[123] Economic Commission for Latin America and the Caribbean (ECLAC/CEPAL) 2013.

[124] Economic Commission for Latin America and the Caribbean (ECLAC/CEPAL) 2010.

[125] Economic Commission for Latin America and the Caribbean (ECLAC/CEPAL) 2010.

the academic analysis of data from a gender perspective."[126] In 2006 through shared efforts by Brazil's IBGE and ECLAC's Social Statistics Unit, members of the Rio Group on Poverty Statistics published a Compendium of best practices in poverty measurement. The Compendium offers a "menu of poverty measurement approaches and methodologies."[127]

Regional Interdisciplinary Network of Innovation Involving Both NSOs and Other Actors

Multi-stakeholder approaches will be key to strengthen NSOs, identify regional priorities, and to ensure trust and legitimacy towards citizens and partners through Big Data. Early coordination of actors working with Big Data across Latin America has been initiated as part of Data-Pop Alliance's Data Space Latin America. The Data Space represents a collective of actors and activities in the Latin American data and development ecosystem working on research, training and advocacy efforts involving Big Data and the SDGs. The Data Space acts as a connecting and sounding board for its members to catalyze and coordinate efforts to maximize their potential around common objectives. CEPEI also is working in coordination efforts on Big Data in the region through its Collaborative Effort on National Data Ecosystem, which supports data-driven decision-making among private and public actors. To achieve this, it promotes the exchange of information between journalists, data scientists, academics, policy makers, and the international community on the implementation, achievements, and limitations of the post-2015 development agenda.

These five trends present opportunities for NSOs and other actors working in the Big Data ecosystem to build on existing frameworks and movements in the region.

4.2 Towards a Regional Multi-Partner Roadmap for Big Data: Building on Existing Regional Strengths and Opportunities

The following recommendations form the basis for a regional, multi-stakeholder roadmap on Big Data in Latin America, describing how NSOs and other regional actors in the Latin American data ecosystem can build on specific existing regional strengths and opportunities to leverage Big Data for official statistics and the SDGs:

1. Creating structures to foster the development and coordination of new and existing projects on Big Data;
2. Mobilizing political awareness and will to ensure policy creation on Big Data; and
3. Developing mechanisms and tools for Big Data use through feedback and learning. Figure 6 below further details recommendations emerging from each of the ongoing regional trends:

[126] Instituto Nacional de Estadística y Geografía (INEGI) de México 2015.

[127] Expert Group on Poverty Statistics 2006.

the Open Data movement, public-private partnerships, regional working groups, emerging best practices for statistics, and an emerging network on Big Data and development in LAC regions.

Create structures to Foster the Development and Coordination of New and Existing Projects on Big Data

As noted earlier, the emergence of public-private partnerships as new forms of collaboration in Latin America allow for an exchange of knowledge and skills between NSOs engaging with Big Data and private and public sector organization providing their technical and computing capabilities. However, working in such partnerships where data is owned outside of NSOs requires additional levels of mediation and negotiation. While NSOs clearly benefit from these exchanges, the value for private sector companies largely reflects their own economic incentives, which can lead to a need for greater compromises that may ultimately mitigate the benefits for NSOs to participate. Further, due to privacy and security concerns, the data governance rules surrounding data use can be both limiting and costly. For example, Telefónica's research is usually contractually executed in its Barcelona headquarters due to privacy and confidentiality concerns; this hinders NSOs' access to data.

As these new forms of partnerships foster new Big Data pilots, a lack of coordinated structures and mandates across institutions has resulted in a number of agencies separately assessing or conducting pilots. These weak legal frameworks pose a major obstacle for many NSOs in the region to effectively engage with Big Data. In Colombia, for example, the Ministry of ICT (MINTIC), the Department for Planning (DNP), and DANE are all simultaneously analyzing potential Big Data cases, with MINTIC being in charge of the national strategy on Open Data and DNP in charge of the national strategy on Big Data. However, these simultaneous efforts are not coordinating. This becomes particularly problematic when reaching out to private sector partners and when negotiating agreements. The case of CDR data best illustrates the limits of ad hoc requests and the need for better-aligned processes, public private partnerships, and broader agreements.

What is needed is both a policy-enabling and practitioner-coordinating environment that promotes stronger NSO leadership within partnerships and that incentivizes coordination among stakeholders across the Latin American Big Data ecosystem. For the SDG process, as well as for any approach related to Big Data, it will be essential for NSOs in the region to establish coordinated efforts in partnership with relevant local stakeholders. In addition to Data-Pop Alliance's Data Spaces, several other actors have been influential in cultivating the Latin American Big Data ecosystem: CEPAL's ongoing work in region connects organizations from different sectors; Fundación Telefónica's survey work on transformations in telecommunications and Internet-based services (in partnership with CAF and ECLAC);[128] and CEPEI's ongoing coordination efforts. These efforts continue to promote coordination among academic, private sector, and public sector stakeholders across Latin America. Additional

[128] Katz 2015.

Figure 6: Roadmap of Recommendations for Incorporating Big Data into NSO projects in LAC

	Create structures to foster the development and coordination of new and existing projects on Big Data	Mobilize political awareness and will to ensure policy creation on Big Data	Develop mechanisms and tools for Big Data use through feedback and learning
Latin American experience of the Open Data movement	Build on corresponding civic engagement pathways	Build on the momentum from the Open Government and Open Data movements as catalysts toward Big Data	Adapt existing data frameworks and standards from the Open Data movement toward Big Data
The emergence of public-private partnerships on Big Data	Develop innovative approaches	Promote and incentivize private sector involvement	Develop frameworks and procedures for ethical data and IP sharing
The presence of strong international and region-wide committees, institutions and working groups (e.g. SDGs)	Drive funding towards the development of Big Data projects	Drive regional coordination and strategy for adopting Big Data policies	Devise big data-based indicators to track progress in regional and global projects
The development of adaptable best practices for NSOs	Field new ideas and "what works" for NSOs	Encourage NSO self assessments through comparative analysis on statistical data collection and production	Share and devise NSO best practices checklists
Regional, coordinated network of innovation involving both NSO and other actors	Coordinate multi-stakeholder engagement and partnerships across Big Data ecosystem	Encourage spectrum of NSO participation in Big Data ecosystem	Develop and maintain platforms for exchanging data between stakeholders

recommendations towards creating structures to foster new and existing Big Data projects include:

1. Build on corresponding civic engagement pathways emerging from the Open Data movement (including visualization tools, APIs, etc) for Big Data;
2. Develop innovative approaches in forming partnerships with private and public sector entities;
3. Drive funding towards the development of Big Data projects through regional committees and working groups;
4. Field new ideas and "what works" among NSOs;
5. Coordinate multi-stakeholder engagement and partnerships across the Big Data ecosystem by fostering regional data ecosystems around key actors and activities to link grassroots groups and startups with large corporations, universities and civil society.

Mobilize political awareness and will to ensure policy creation on Big Data

As noted above, the Latin American experience with the Open Government and Open Data movement has evoked strong political will across governments in the region. The Open Data movement has fostered the exploration and coordination of non-NSO actors around public data activities; though the impact of open data across the region has been limited, political willingness and support continues to increase. In addition, the existence of regional working groups and institutions such as ECLAC provide international interests and spotlight in the development of the region as well as access to financial resources.

However, the lessons learned from the history of these movements include the need for the development of frameworks and impact assessments at an earlier stage. Much of the recent criticism of the Open Data movement have been the limited impact (and sometimes, civic interest) of costly open data government initiatives initiated by global rallying cries to open data. The policy-first approach neglected the practical considerations necessary to carry out projects in a concrete way and assess their impact.

Regional actors working in Big Data must seek synergies and consider lessons from these parallel data movements in order to mobilize and drive political will and resources toward the creation and development of national Big Data strategies. This includes the coordination of academic and technical experts with advocates and civic hackers towards the development of shared knowledge and frameworks. To guarantee a human-centered and responsible development, there also needs to be constant dialogue with pressure groups, such as Human Rights advocates and journalists. NSOs should thus communicate transparently and openly about Big Data activities and partnerships. This also includes evaluating the capacities of individuals and groups to constructively engage in society through and about data (e.g. data literacy). A greater push towards literacy can enable greater civic participation and demand from their governments toward fostering policy-enabling environments for Big and Open Data.

Additional recommendations towards mobilizing political awareness and will to ensure policy creation on Big Data include:

1. Promote and incentivize private sector involvement, via the organization of data challenges and promotion of financial and in-kind support to local entrepreneurs and startups;
2. Drive coordination and strategy through regional coordinating institutions and working groups for adopting Big Data policies;
3. Capitalize on synergies created by the Global Partnership for Sustainable Development Data;
4. Encourage NSO self assessments through comparative analysis on statistical data collection and production;
5. Encourage spectrum of NSO participation in Big Data ecosystem.

Develop Mechanisms and Tools for Big Data Use through Feedback and Learning

The proliferation of Big Data projects, pilots and actors has generated increasing interest in the potential to solve global issues; however global frameworks and models for addressing the downfalls of Big Data have remained elusive. For example, in considering the ethical and privacy concerns related to the reidentifcation of personally identifiable information (PII), there are considerable gaps in understanding the nature of responsible data use and the development of corresponding legal frameworks. Additionally, as many governments and other actors consider the use of algorithmic methodologies toward data-driven policymaking, knowledge-sharing on how to address the implications of these methodologies also remain largely unexplored.

What is needed are mechanisms and tools for Big Data use towards greater knowledge sharing and coordination among stakeholders. This is particularly for NSOs in developing regions such as Latin America where limited resources create less room for experimentation. NSOs by mandate both collect, coordinate and disseminate data for government agencies and other actors on society. However, the proliferation of new data sources via Big Data has made this task of dissemination increasingly complex through the lack of formats and standards, the sheer volume data and the nature of the data collection process. NSOs used to oversee the data collection process; now they are picking up crumbs from data sources where the data collection occurs upstream from them.

As NSOs experiment with Big Data, lessons learned in facilitating the task of dissemination would be valuable towards the development of best practices among NSOs and decrease the barriers of entry for other NSOs to convert and interact with new data sources.

As stated earlier, in terms of the creation of frameworks for data protection, many Latin American countries have Omnibus Data Protection laws similar to EU data protection laws, where citizens are able to control the use of their personal data held by public or private entities.

Additional recommendations towards developing mechanisms and tools for Big Data use include:

1. Adapt existing data frameworks and standards from the Open Data movement toward Big Data.
2. Develop frameworks and procedures for ethical data and IP sharing, possibly with an Ethics board committee in each NSO.
3. Devise big data-based indicators to track progress in regional and global projects.
4. Share and devise NSO best practices checklists.
5. Develop and maintain platforms from exchanging data between stakeholders.

While Latin American NSOs will continue to play a pivotal role in the evolution of official statistics and the achievement of the SDGs in the region, this report has highlighted that they are in fact not alone in these efforts and must coordinate and work with other actors—government agencies, international organizations, civil society, universities, etc.—in order to realize the full potential of Big Data for official statistics and the SDGs. For NSOs in the LAC region, it will be just as important to be aware of and engage with the wider ecosystem—also with regard to the SDG process.

Glossary

Terms

- **Big Data** = The ecosystem created by the concomitant emergence of 'the 3 Cs of Big Data': Digital Crumbs—pieces of data passively emitted and/or collected by digital devices which constitute very large data sets and streams and contain unique insights about their behaviors and beliefs; Big Data Capacities—what has also been referred to as Big Data Analytics, that is the set of tools and methods, hardware and software, know-how and skills, necessary to process and analyze these new kinds of data—including visualization techniques, statistical machine-learning and algorithms, etc; Big Data Communities—which describe the various actors involved in the Big Data ecosystem, from the generators of data to their analysts and end-user—i.e. potentially the whole population.

- **big data** = refers to the 1st C of Big Data (above), i.e. the streams and sets resulting from humans leaving digital traces when using cell-phones (call detail records), credit cards (transactions), transportation (subway or bus records, EZ pass logs), social media and search engines, or having their actions picked up by sensors, whether physical (electrical meters, weigh sensors on a truck) or remote (satellites, cameras).

- **crowdsourcing** = the practice, usually led through digital platforms (SMS, internet, etc), of enlisting a large number of people to contribute to a particular task or effort.

- **exhaust data** = data that are passively emitted from cell phones, sensors, social media and other platforms as digital translations of human actions and interactions.

- **thick data** = qualitative information that provides insight into the emotional aspects of human behavior, as opposed to thin data which mainly focuses on quantitative information which provides less robust insight into qualitative aspects of the observed behaviors.

- **webscraping** = A computer software technique for automating the extraction of information from websites.

Acronyms

- **CDP** = CDP Worldwide, a company that reports climate change, water, supply chain, forest, and other environmental data, with the goal to prevent climate change and protect the environment.

- **CDR** = Call Detail Record, The technical name for mobile phone data recorded by all telecom operators. CDRs contain information about the locations of those sending and receiving calls or text messages through operators' networks, as well as data on time and duration.

- **ECLAC/CEPAL** = Economic Commission for Latin America and the Caribbean (in Spanish, *Comisión Económica para América Latina*), a regional commission of the UN aimed at promoting economic development in the region.

- **EU** = European Union.

- **FBK** = Fondazione Bruno Kessler, is a private entity charged with keeping the province of Trento, Italy, in the mainstream of European and international research.

- **HADOOP** = A system for maintaining a distributed file system that supports the storage of large-scale (Terabytes or Petabytes of content), and the parallel processing of algorithms against large data collections, which requires a programming language such as Java or Python.

- **HDX** = Humanitarian Data Exchange, a platform for sharing data whose goal is to make data easy to find and use.

- **HHI** = Harvard Humanitarian Initiative, an interdisciplinary research center at Harvard University that specializes in humanitarian relief and crisis response.

- **ICT** = Information and communications technology, which pertains to the convergence of audio-visual and telephone networks with computer networks through a single cabling or link system, and the economic and infrastructural implications of such tendencies.

- **IEAG** = Independent Expert Advisory Group, a branch of the UN Secretary-General that has been asked to make recommendations for how shape "an ambitious and achievable vision" for a future development agenda beyond 2015 to succeed the United Nations Millennium Development Goals.

- **NSO** = National Statistics Office, a leading statistical agency in a national statistical system.

- **OCHA** = Office for the Coordination of Humanitarian Affairs, a part of the UN aimed at improving emergency responses, which includes developing the Humanitarian Data Exchange.

- **ODI** = Overseas Development Institute, independent think tank on international development and humanitarian issues, based in the UK.

- **OSILAC** = Observatory for the Information Society in Latin America and the Caribbean, whose objective is to improve Information and Communications Technologies (ICT) statistics in Latin America.

- **RIVAF** = Rapid Impact and Vulnerability Analysis Fund, a UN project wherein the UNODC (UN Office on Drugs and Crime) and UN Global Pulse (see below) researched how crises may impact crime levels.

- **SDG** = Sustainable Development Goals, established by the UN Division for Sustainable Development to promote and coordinate implementation of the sustainable development agenda of the United Nations.

- **UN** = United Nations.
- **UN Global Pulse** = An initiative by the UN to use big data for development and humanitarian action, which consists of a group of data innovation projects about a range of global issues.
- **UNDOC** = United Nations Office on Drugs and Crime, which conducts field-based projects to fight illegal drugs and crime, as well as research.
- **UNECE** = United Nations Economic Commission for Europe, whose objective is to promote European economic integration.
- **UNFPA** = United Nations Population Fund, which works to promote safe pregnancy and healthy childbirth.
- **UNSD** = United Nations Statistics Division, which collects and reports economic and social statistics.

Annexes

Annex 1: Taxonomy and Examples of Big Data Sources

Types	Examples	Opportunities
	Category 1: Exhaust data	
Mobile-based	Call Details Records (CDRs) GPS (Fleet tracking, Bus AVL)	Estimate population distribution and socioe-conomic status in places as diverse as the U.K. and Rwanda
Financial transactions	Electronic ID E-licenses (e.g. insurance) Transportation cards (including airplane fidelity cards) Credit/debit cards	Provide critical information on population movements and behavioral response after a disaster
Transportation	GPS (Fleet tracking, Bus AVL) EZ passes	Provide early assessment of damage caused by hurricanes and earthquakes
Online traces	Cookies IP addresses	Mitigate impacts of infectious diseases through more timely monitoring using access logs from the online encyclopedia Wikipedia
	Category 2: Digital Content	
Social media	Tweets (Twitter API) Check-ins (Foursquare) Facebook content YouTube videos	Provide early warning on threats ranging from disease outbreaks to food insecurity
Crowd-sourced/ online content	Mapping (Open Street Map, Google Maps, Yelp) Monitoring/Reporting (uReport)	Empower volunteers to add ground-level data that are useful notably for verification purpose
	Category 3: Sensing data	
Physical	Smart meters Speed/weight trackers USGS seismometers	Sensors have been used to assess the demand for using sensors to estimate demand for high efficiency cook-stoves at different price points in Uganda or willingness to pay for chlorine dispensers in Kenya
Remote	Satellite imagery (NASA TRMM, LandSat) Unmanned Aerial Vehicles (UAVs)	Satellite images revealing changes in, for example, soil quality or water availability have been used to inform agricultural interventions in developing countries

Annex 2: Uses of Big Data Towards Monitoring the SDGs

SDGs adopted by the OWG	Big data examples	What is monitored	How is monitored	Country(ies)	Year	Advantages of using big data
1. End poverty in all its forms everywhere	Satellite data to estimate poverty[129]	Poverty	Satellite images, night-lights	Global map	2009	International comparable data, which can be updated more frequently
	Estimating poverty maps with cell-phone records[130]	Poverty	Cell phone records	Côte d'Ivoire	2013-14	
	Internet-based data to estimate consumer price index and poverty rates[131]	Price indexes	Online prices at retailers websites	Argentina	2013	Cheaper data available at higher frequencies
	Cell-phone records to predict socio-economic levels	Socio-economic levels	Cell phone records	"Major city in Latin America" (Actually Mexico-City)	2011	Data available more regularly and cheaper than official data; informal economy better reflected
2. End hunger, achieve food security and improved nutrition and promote sustainable agriculture	Mining Indonesian Tweets to understand food price crises[132]	Food price crises	Tweets	Indonesia	2014	
	Uses indicators derived from mobile phone data as a proxy for food security indicators[133]	Food security	Cell phone data and airtime credit purchases	A country in Central Africa	2014	
	Use of remote-sensing data for drought assessment and monitoring	Drought	Remote sensing	Afghanistan, India, Pakistan[134]	2004	
				China[135]	2008	
3. Ensure healthy lives and promote well-being for all at all ages	Internet-based data to identify influenza breakouts[136]	Influenza	Google search queries	US	2009	Real-time data; captures disease cases not officially recorded; data available earlier than official data
	Data from online searches to monitor influenza epidemics[137]	Influenza	Online searches data	China	2013	
	Detecting influenza epidemics using Twitter	Influenza	Twitter	Japan	2011	
	Monitoring influenza outbreaks using Twitter	Influenza	Twitter	US	2013	
	Systems to monitor the activity of influenza-like-illness with the aid of volunteers via the internet	Influenza	Voluntary reporting through the internet	Belgium, Italy, Netherlands, Portugal, United Kingdom, United States	ongoing	

[129] Christopher D. Elvidge, Sutton, et al. 2009.

[130] Smith-Clarke, Christopher and Mashhadi, Afra and Capra, Licia 2014.

[131] Cavallo 2013.

[132] United Nations Global Pulse 2014.

[133] Soto et al. 2011.

[134] Thenkabail, Gamage, and Smakhtin 2004.

[135] Zhang et al. 2008.

[136] Ginsberg et al. 2009.

[137] Yuan and al. 2013.

(continued)

SDGs adopted by the OWG	Big data examples	What is monitored	How is monitored	Country(ies)	Year	Advantages of using big data
	Cell-phone data to model malaria spread	Malaria	Cell-phone data	Kenya	2012	
	Using social and news media to monitor cholera outbreaks	Cholera	Social and news media	Haiti	2012	
	Google dengue trends	Dengue	Web search queries	Argentina, Bolivia, Brazil, India, Indonesia, Mexico, Philippines, Singapore, Thailand, Venezuela	ongoing	
	Monitoring vaccine concerns to help tailor immunization programs	Vaccine concerns	media reports (e.g., online articles, blogs, government reports)	144 countries	2013	Data not available otherwise; expensive to collect data through survey
	Monitoring vaccine concerns	Vaccine concerns	Twitter	US	2011	
	Analysis of Twitter used to track HIV incidence and drug-related behaviors	HIV, drugs use	Twitter	US	2014	
7. Ensure access to affordable, reliable, sustainable and modern energy for all	Satellite data to estimate electric power consumption[138]	Electric power consumption	Satellite images	21 countries	1997	Regular updates
8. Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all	Light emissions picked up by satellites to estimate GDP growth[139]	GDP growth	Satellite images	30 countries	2012	Informal economy better reflected; information available at sub-national level; improves estimates for countries with poor national accounts data
	Using night-lights to estimate GDP at sub-national levels[140]	GDP at sub-national levels	Satellite images	China, India, Turkey, US	2007	
	Internet-based data to monitor inflation in real time	Inflation	Prices from online retailers	Argentina, Brazil, Chile, Colombia, Venezuela	2012	Cheaper data available at higher frequencies

[138] C. D. Elvidge et al. 1997.

[139] Henderson, Storeygard, and Weil 2012.

[140] Sutton, Christopher D. Elvidge, and Ghosh 2007.

(continued)

SDGs adopted by the OWG	Big data examples	What is monitored	How is monitored	Country(ies)	Year	Advantages of using big data
9. Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation	Map showing internet devices which could be logged in using default passwords or no passwords. Despite biases towards unsecure devices, the map may reflect online usage around the world[141]	Map with internet devices by location	Internet tools to scan all addresses of the fourth version of the internet protocol	World	2012	Easier, cheaper, quicker than internet use surveys. Disadvantages: illegal and may not be able to be reproduced with the newest internet protocols
10. Reduce inequality within and among countries	Mapping socio-economic status by analyzing airtime credit and mobile phone datasets[142]	Wealth and inequality	Airtime credit purchases	Côte d'Ivoire	2013	Disadvantage: no ground truth data to compare it with (last censuses unreliable)
11. Make cities and human settlements inclusive, safe, resilient and sustainable	Light emissions picked up by satellites to estimate urban extent	Urban extent	Satellite images	Global	2005	Globally consistent way to map urban extent; more regular updates
	Use of data from transport cards to construct a picture of individual journeys and how the bus and train networks are used by the public	Transport use and journeys	Transport cards data	UK		More detailed and more frequent than survey data
	Times series of satellite images of flooded areas are used to identify flood risk areas	Flood hazard and risk	Satellite images	Namibia	2014	Data available frequently
	Analysis of the temporal evolution of nightlights along the river network to obtain a global map of human exposure to floods	Night lights as a proxy for population/infrastructure along the river network	Satellite images	Global	1992-2012	
	Using satellite imagery, GIS and precipitation data to produce a flood risk map along the Niger-Benue river	Flood risk	Satellite images	Nigeria	2014	
	Using satellite remote sensing and GIS techniques for flood hazard and risk assessment in Chamoli district, Uttarakhand, India	Flood hazard and risk	Satellite images	India	2014	
	Assessing flood impact with cell phone records	Flood impact	Cell phone records	Mexico	2014	
	Analysis of Twitter data during hurricane Sandy to identify which data may be useful in disaster response[143]	Tweets about the hurricane	Twitter	US	2012	

[141] Carna Botnet 2012.

[142] Gutierrez, Krings, and Blondel 2013.

[143] Statistics Without Borders and Humanity Road 2013.

(continued)

SDGs adopted by the OWG	Big data examples	What is monitored	How is monitored	Country(ies)	Year	Advantages of using big data
13. Take urgent action to combat climate change and its impacts	Satellite scan to monitor population and energy related greenhouse gas emissions[144]					Separate emissions of urban populations from other sources; more regular updates
	Satellite images to measure net primary production[145]					Regular updates
	Methane observations made from space combined with Earth-based remote sensing column measurements[146][147]	Methane	Satellite measurements	US	2014	
16. Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels	Use of mobile phone and demographic data to predict crime in London[148]	Crime	Mobile phone and demographic data	UK		
	Using the 'Global Data on Events, Location and Tone (GDELT)', a news stories dataset, to crunch the numbers of violent events in a conflict[149]	Violent events	News stories database	Syria	2013-14	
Measures beyond GDP	Cell-phone records to predict socio-economic levels[150]					Data available more regularly and cheaper than official data; informal economy better reflected

[144] Christopher D. Elvidge, Baugh, et al. 1997.

[145] Net primary production refers to how much carbon dioxide vegetation takes in during photosynthesis minus how much carbon dioxide the plants release during respiration.

[146] Kort et al. 2014.

[147] Schneising et al. 2014.

[148] Bogomolov et al. 2014.

[149] Earl et al. 2004.

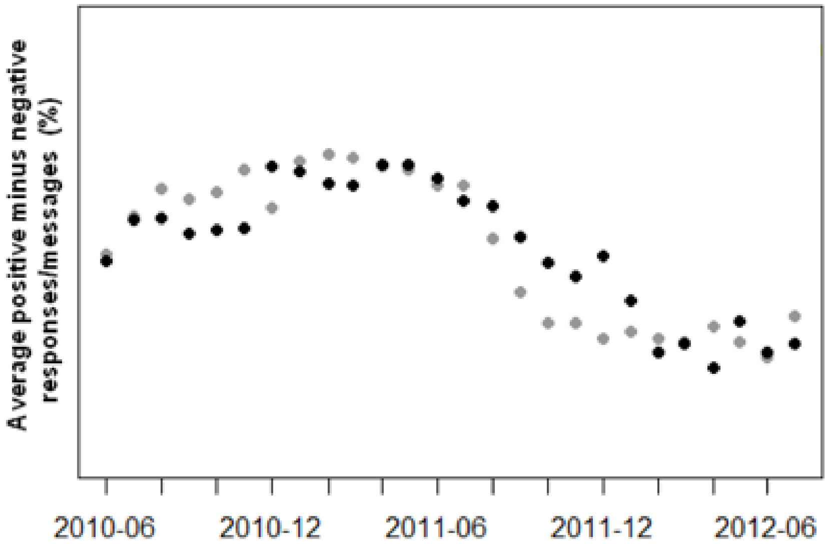
[150] Soto et al. 2011.

Annex 3: Analysis of Social Media Messages by Statistics Netherlands

The Statistics Netherlands studied publicly available social media messages created on various social media platforms, such as Twitter and Facebook, as well as the public messages posted on news sites, web forums and blogs. The messages were obtained from a commercial company that routinely harvested all publicly available messages written in Dutch on the Dutch-language part of the web.

Both the content and the sentiment of the messages were studied. Studies of the content of messages in Dutch on Twitter, the social media platform on which most publicly available Dutch-language messages are created, revealed that nearly 50% of those messages were composed of "pointless babble." The remainder predominantly discussed spare-time activities (10%), work (7%), media (television and radio) (5%) and politics (3%). Use of these more serious messages was hampered by the less serious "babble" messages.

Figure: Dutch consumer confidence (grey) and the overall sentiment in Dutch social media messages on a monthly basis (black). Dutch articles are used as search terms. The social media sentiments in December is considerably more positive compared to the sentiment in the months before and after.



Daas, Piet J.H. and Puts, Marco J. and Buelens, Bart and van den Hurk, Paul A.M. "Big Data and Official Statistics." In: *New Techniques and Technologies for Statistics (NTTS)*. 2013. URL: http://ec.europa.eu/eurostat/cros/content/ntts-2013_en

Determination of the sentiment in all messages created on all available platforms revealed a highly interesting potential use of these data for statistics. With a query language and a web interface, messages were selected from the database. Messages were classified in positive, negative and neutral. The sentiment in these messages was found to be highly correlated with Netherlands consumer confidence, in particular with sentiments regarding the economic situation. A consumer confidence index is produced every month by Statistics Netherlands using survey data from a random sample from the population register. While social media messages are generated by 70% of the Dutch population. The latter relation was stable on a monthly and on a weekly basis. Daily figures, however, displayed highly volatile behavior suggesting that it is possible to produce monthly and weekly sentiment indicators comparable with consumer confidence. The latter indicators can be produced on the first working day following the week studied, demonstrating the ability to deliver results quickly. Only in December the numbers did not relate, where a much more positive sentiment occurred in social medial, removing all messages including words for Christmas and New Year's Eve reduced these peaks.[151]

Annex 4: Price Indices by Cavallo (MIT)

The objective of this project was to investigate and show how scraping the web for online prices could provide real-time insights on price dynamics. Prices collected from online retailers can be used to construct high-frequency price indexes that complement official statistics. At MIT a research team around Alberto Cavallo used data that were collected between October 2007 and March 2011 from the largest online supermarkets in five Latin American countries and studied their ability to match official inflation estimates. The data is collected in Argentina, Brazil, Chile, Colombia, Uruguay and Venezuela using a scraping software that records, on a daily basis, the price of bread sold or advertised in online supermarkets. Then the daily inflation rate of bread for each country is calculated. They focused on Argentina, where official statistics have been criticized in recent years. The data Online price indexes approximate both the level and main dynamics of official inflation in Brazil, Chile, Colombia, and Venezuela. By contrast, Argentina's online annual inflation rate is consistently two to three times higher than in official estimates.

Partners: PriceStats and the Billion Prices Project at MIT Argentina, Brazil, Uruguay, Venezuela.[152]

Annex 5: Luminosity Data as a Proxy for Economic Statistics

A pervasive issue in social and environmental research has been how to improve the quality of socio-economic data in developing countries. Given the shortcomings of standard sources, the

[151] United Nations Statistical Commission 2014a.

[152] United Nations Global Pulse 2011.

present study examines luminosity (measures of nighttime lights visible from space) as a proxy for standard measures of output (gross domestic product). The researchers compare output and luminosity at the country level and at the latitude and longitude grid-cell level for the period 1992-2008. They find that luminosity has informational value for countries with low-quality statistical systems, particularly for those countries with no recent population or economic censuses.[153]

Annex 6: An Information System for Prices in Agriculture

SIPSA provides information for agriculture prices. The information is based on wholesale prices of food, food supply to cities and inputs and factors associated with agricultural production and livestock.

The first information system aims to collect information on wholesale prices in the moment prices are formed. This information is collected via the SIPSA app or web interfaces and disseminated through daily, weekly and monthly newsletters. Each component seeks to meet different information needs. The regional daily newsletter is aimed primarily at those persons present in the markets, to give them evidence and negotiating tools when conducting transactions. The national daily bulletin shows the behavior of prices in seven major cities. This is a tool especially for all those involved decision makers, both public and private. The weekly newsletter explains the different events that affected the marketing of agricultural products throughout the week. The monthly newsletter shows the aggregate behavior of wholesale prices compared with the immediately preceding month. The other two components, food supply factors associated inputs and agricultural production, register the quantities of agricultural products entering and leaving the cities at major markets in the country; as well as the retail price of the main inputs and factors associated with the agricultural production and livestock in the country.

Source: PARIS21

Annex 7: Open Intelligence in Mexico

The Mexican startup Open Intelligence develops cloud-based analytics and communication platforms that support governments and other public sector institutions to make policy decisions based on data. OPI developed a comprehensive public data-warehouse and centralizes millions of data points about social and economic trends in Mexico and provides corresponding visualization and analytics through its platform. For clients aiming to generate more topic-specific data, OPI's associated mobile apps facilitate on the ground data collection [154].

[153] Chen and Nordhaus 2011.

[154] For example: `http://brujulacd.mx/`

The Mexican Ministry of Interior for example used their platform to understand neighborhood crime rates. The Ministry analyzed the relationship between education, single parent families, and over one thousand other variables. Later that year, the ministry started redesigning its policy and resource allocation strategies based on OPI's contributions.[155]

Annex 8: Leveraging Big Data Sources and Techniques Based on CDRs to Analyze Socio-economic Outcomes and Processes in Colombia: The Cases of Public Safety and Social Development

The first of two pilot studies being conducted by Data-Pop Alliance, Telefónica, and Bruno Kessler Foundation, with funding from The World Bank, focuses on public safety and crime in Bogotá. The research looks into crime data, obtained from the Policía Nacional de Colombia (Colombia's national police), in conjunction with other types of data, primarily call detail record (CDR) data from 2014 provided by Telefónica. The objective is to see how alternate data sources can help understand and predict the emergence of crime hotspots, both to predict/prevent future crime, and to understand what characterizes areas where crime rates are particularly high or particularly low.

Generally speaking, one of the best predictors of future crime is past crime, meaning that the National Police's crime data on past crime reports could, in and of itself, provide valuable insight into where future crime will occur. However, the value of bringing in outside data, such as data on cellphone calls and SMS patterns in Bogotá, lies in the fact that they provide provide insights that are lost when relying on crime data alone.

Indeed, possible correlates of crime, such as income and social networks, will be implicitly recorded in the past crime data that is used for prediction, but those patterns emerge more readily when the crime data is intersected with data from other sources. Furthermore, it is useful to develop alternate methods of crime prediction that allow NSOs to monitor and predict crime even when data on previous crimes is not available or reliable. Thus, these pilots aim to use CDRs and other data sources to build a more contemporary, complete, and complex picture of social outcomes and underlying processes in Latin America in general and Colombia.

The second pilot study focuses on social outcomes. The research will use CDR data from 2014 provided by Telefónica to derive socio-economic indicators for Bogotá, focusing in particular on poverty and social cohesion. Context data will be taken from the 2014 Encuesta Multipropósito (Multipurpose survey) of Bogotá, carried out by the city's Secretaría Distrital de Planeación (SDP) in partnership with DANE, which gathered microdata from a representative sample of households in each of 19 localities of the city and 31 of its surrounding municipalities. The data cover 14 topics, including household living conditions, health, education levels, and household spending. This survey data will be aggregated to a higher geographic level to preserve household

[155] GSMA Intelligence n.d.

anonymity, and meaningful features will be analyzed by Data-Pop Alliance's research team and affiliates who have previous experience working with data from DANE. This data can serve as ground-truth or help refine the predictive power in the analysis of the CDR data, as needed.

As with the first pilot study on crime, the ultimate objective of this program is to help Colombia's Departamento Administrativo Nacional de Estadística (DANE) explore whether and how Big Data sources and techniques—and specifically as they relate to CDRs—can be leveraged to derive social indicators in ways that could be incorporated into the country's official statistics workflow. For each pilot study, the expected outputs are:

- an empirical research paper of publishable quality in top academic journals, written by the researchers from Data-Pop Alliance, Telefónica, and Bruno Kessler Foundation;
- a version of the previous paper, adapted by Data-Pop Alliance, presenting the key lessons and findings, with accompanying codes and visuals, for use by DANE;
- a repository of the codes and visualization tools used in the project, published under a Creative Commons license and intended for reuse as a learning and training tool.

Annex 9: Major Information Technology Components

Apache Hadoop. A system for maintaining a distributed file system that supports the storage of large-scale (Terabytes or Petabytes of content), and the parallel processing of algorithms against large data collections, which requires a programming language such as Java or Python.

Apache Spark. A fast and general purpose engine for large-scale data processing that works in support of Hadoop or in-memory databases. Requires a programming language such as Java or Python.

Java programming language. A general purpose systems engineering language that supports the creation of efficient algorithms for data analysis.

Pig and Hive as programming tools for data manipulating (i.e. to query data on Hadoop clusters) before using statistical software (R, SAS, SPSS or similar).[156]

Python programming language. A general purpose systems engineering language that supports rapid prototyping and efficient algorithms for data analysis.[157]

R, PostgreSQL or Weka as open source and free technologies to analyze Social Media content

[156] Eurostat 2014.

[157] American Association for Public Opinion Research (AAPOR) 2015.

such as Twitter.

Annex 10: Aggregated SWOT analysis for LAC NSOs and Big Data

Strengths	**Weaknesses**
Region is getting more interest from private sector. Mobile technology, Internet and social media are widely available (though there are gaps between rural and urban areas). In many LAC countries, statistical systems have a solid, long tradition of census and surveys. NSIs by mandate and design are trained and prepared to work with data (both in terms of technical facilities and legislation). NSIs have a well-established process for monitoring the MDGs (i.e. virtual training, exchange on regional level). There is a presence of strong region-wide committees and institutions and working groups such as ECLAC. Universities are increasingly starting master's programs on data science. First big data pilots and roll-outs have been realized. There are several examples of usage of Big Data in the region, initiated by other actors who could become potential partners. A lot of data sources are available—LAC is seeing a digital revolution. There is technological restructuring (towards GSBPM) at some NSIs; most are in the process of building up data warehouses; many work with Hadoop. Many are in the process to switch to SDMX.	In LAC region there is little culture of "evidence based decision making." Overall research and innovation culture in the region is weak. Weak legal frameworks limit many NSIs in the region. There is limited or bad inter-operationality between different agencies. Institutions have little knowledge about concept of Big Data. Openness is proclaimed but not yet adhered to. Homepages and other distributions channels are still very weak, weak engagement with data beneficiaries. Big Data is not a high priority. Big Data efforts are not embedded in a larger strategy. Progress of pilots depends on goodwill and engagement of champions. There is a lack of human capacity to work with data; data literacy. There are limited opportunities to get outside support, i.e. via summer interns. There are limited opportunities for partnerships and collaboration. Efforts are parallel instead of merging related areas, like SGDs, open data, big data, innovation There is no culture of Public Private Partnerships. There is little exchange other with external stakeholders, i.e. Start-up, CSOs.

Opportunities	Threats
Process of SGDs could strengthen mandate of NSIs in the region. First big data best practice examples within the region can be learned from. There is a vibrant movement on Open Data and many Open Source proponents. Vibrant discussions on Internet Government take place across the region (see Marco Civil as best practice example for multi stakeholder processes). "Champions" in the organizations are eager to work with and discuss Big Data. Vibrant technology ecosystems exist in many countries in LAC. There is demand for data spaces across the region that could be initiated by NSIs. Ideas and experiences can be exchanged with other NSIs, also at the UN level (however this could be a risk for those who are not included); a Community of practices could be established. A platform for knowledge sharing could be built; i.e. wiki for Big Data + NSIs in the LAC region.	Other agencies could occupy the topic. There is little debate about potential risks and negative implications of Big Data. There is confusion about Big Data as an ecosystem and Big Data as a data source, and between Big Data and other data sources, i.e. open data. Reinventing the wheel—NSIs face competition instead of collaboration (both between agencies, as well as between countries) There is no sustainable access to certain data sets, i.e. CDR data; plus already very unsustainable way of exchanging / accessing data, i.e. administrative data, not based on legacies Organizations could copy instead of invent according to local needs Budgets in some countries (Mexico, Colombia etc.) could be cut due to sinking oil prices. NSIs remain "closed shops."

Bibliography

American Association for Public Opinion Research (AAPOR). *AAPOR Report on Big Data*. 2015. URL: https://www.aapor.org/AAPOR_Main/media/Task-Force-Reports/BigDataTaskForceReport_FINAL_2_12_15_b.pdf.

Ballivian, Amparo and William Hoffman. *Public-Private Partnerships for Data*. 2015. URL: http://data.worldbank.org/sites/default/files/issue-paper-financing-the-data-revolution-ppps_0.pdf.

Barbosa, Ariel. *Global Information Society Watch 2014: Communications surveillance in the digital age, Colombia*. 2014. URL: https://www.giswatch.org/sites/default/files/hacking_information_on_the_peace_talks_in_colombia.pdf.

Bibolini, Lucia and Henry Lancaster. *2014 Latin America – Telecoms, Mobile and Broadband Overview*. 2014. URL: http://www.budde.com.au/Research/2014-Latin-America-Telecoms-Mobile-and-Broadband-Overview.html?r=51.

Bogomolov, Andrey et al. *Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data*. 2014. DOI: arXiv:1409.2983. URL: http://arxiv.org/abs/1409.2983.

boyd, danah and Kate Crawford. "Critical Questions for Big Data." In: *Information, Communication & Society* 15.5 (2012). ISSN: 1468-4462. DOI: 10.1080/1369118X.2012.678878. URL: http://www.tandfonline.com/doi/pdf/10.1080/1369118x.2012.678878.

Campos, Guilherme. *Where Does Latin America Stand in Terms of Big Data Adoption?* 2015. URL: http://www.nearshoreamericas.com/latin-america-stand-terms-big-data-adoption/.

Carna Botnet. "Internet Census 2012." In: (2012). URL: http://internetcensus2012.bitbucket.org/paper.html.

Cavallo, Alberto. "Online and official price indexes: Measuring Argentina's inflation." In: *Journal of Monetary Economics* 60.2 (2013), pp. 152–165. ISSN: 0304-3932. DOI: 10.1016/j.jmoneco.2012.10.002. URL: http://www.sciencedirect.com/science/article/pii/S0304393212000967.

Cavenaghi, Suzana. *Data Revolution: Is Latin America prepared and ready to engage?* 2015. URL: http://paa2015.princeton.edu/uploads/153763.

CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS). "Cracking patterns in big data saves Colombian rice farmers huge losses." In: *2014 Annual Report* (2014). URL: https://ccafs.cgiar.org/research/annual-report/2014/cracking-patterns-in-big-data-saves-colombian-rice-farmers-huge-losses.

Chen, Xi and William D. Nordhaus. "Using luminosity data as a proxy for economic statistics." In: *Proceedings of the National Academy of Sciences* 108.21 (2011), pp. 8589–8594. DOI:

10.1073/pnas.1017031108. URL: http://www.pnas.org/content/108/21/8589.abstract.

CIVICUS. *The Data Shift*. URL: http://civicus.org/thedatashift/.

Clark, Liat. "Nuria Oliver: what big data and the Mexican pandemic taught us." In: *Wired UK* (2013). URL: http://www.wired.co.uk/news/archive/2013-10/17/nuria-oliver.

Cobos, María Isabel, Tim Miller, and Magda Ruiz Salguero. "Hacia la armonización de las estimaciones de mortalidad materna en América Latina: hallazgos de un estudio piloto en ocho países." In: *Naciones Unidas, Santiago, Chile*. 108th ser. (2013). ISSN: 1680-899. URL: http://repositorio.cepal.org/bitstream/handle/11362/7143/LCL3735_es.pdf?sequence=1.

Congreso de la República de Colombia. *Ley Estatutaria 1266 de 2008*. 2008. URL: http://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=34488.

— *Ley Estatutaria 1581 de 2012*. 2012. URL: http://www.secretariasenado.gov.co/senado/basedoc/ley_1581_2012.html.

— *Proyecto de ley 1753 de 2015 cámara por la cual se expide el Plan Nacional de Desarrollo 2014-2018 'Todos Por un Nuevo País'*. 2014. URL: https://colaboracion.dnp.gov.co/CDT/Prensa/ArticuladoVF.pdf.

Cordero, Arturo Sevilla. *Colombia avanza en una mejor calidad de vida*. 2016. URL: http://docplayer.es/9744214-Colombia-avanza-en-una-mejor-calidad-de-vida.html.

Culzac, Natasha. "Egyptian police 'using social media apps' to trap gay people." In: *The Independent* (2014). URL: http://www.independent.co.uk/news/world/africa/egypts-police-using-social-media-and-apps-like-grindr-to-trap-gay-people-9738515.html.

Daas, Piet J.H. and Puts, Marco J. and Buelens, Bart and van den Hurk, Paul A.M. "Big Data and Official Statistics." In: *New Techniques and Technologies for Statistics (NTTS)*. 2013. URL: http://ec.europa.eu/eurostat/cros/content/ntts-2013_en.

Daas, Piet and Mark van der Loo. *Big data (and official statistics)*. 2013. DOI: 10.2901/Eurostat.C2013.001. URL: http://www.unescap.org/sites/default/files/2-Big%20Data%20(and%20official%20statistics)-Netherlands-presentation.pdf.

Data Revolution for Sustainable Development (IEAG), United Nations Secretary-General's Independent Expert Advisory Group on a. *A World That Counts: Mobilising The Data Revolution for Sustainable Development*. 2014. URL: http://www.undatarevolution.org/report/%20http://www.undatarevolution.org/wp-content/uploads/2014/11/A-World-That-Counts.pdf.

Datafloq. *Cignifi Recognizes Patterns in Mobile Usage For Credit Scores*. 2015. URL: https://datafloq.com/cignifi/.

Digiampietri, Luciano A et al. "BraX-Ray: An X-Ray of the Brazilian Computer Science Graduate Programs." In: *PLoS ONE* 9.4 (2014). ISSN: 1932-6203. DOI:

10.1371/journal.pone.0094541. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3984164/%20http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3984164/pdf/pone.0094541.pdf.

Diniz, Gustavo Macedo, B Girard, and F Perini. *Enabling Openness: The future of the information society in Latin America and the Caribbean*. 2013.

Earl, Jennifer et al. "The Use of Newspaper Data in the Study of Collective Action." In: *Annual Review of Sociology* 30.1 (2004), pp. 65–80. ISSN: 0360-0572. DOI: 10.1146/annurev.soc.30.012703.110603. URL: http://www.annualreviews.org/doi/abs/10.1146/annurev.soc.30.012703.110603.

Economic Commission for Latin America and the Caribbean (ECLAC/CEPAL). *Development of Official Statistics in the Region*. 2010. URL: http://repositorio.cepal.org/bitstream/handle/11362/3146/2010-695_ReportLAC_en.pdf.

— *Consenso de Montevideo sobre Población y Desarollo*. 2013. URL: http://www.cepal.org/celade/noticias/documentosdetrabajo/8/50708/2013-595-consenso_montevideo_pyd.pdf.

— *Big data and open data as sustainability tools*. 2014. URL: http://www.cepal.org/en/publications/37158-big-data-and-open-data-sustainability-tools-working-paper-prepared-economic.

— *Statistical activities in Latin America and the Caribbean: Recent achievements and next challenges*. 2015. URL: https://documents-dds-ny.un.org/doc/UNDOC/GEN/N14/683/08/PDF/N1468308.pdf?OpenElement.

— *The new digital revolution: From the consumer Internet to the industrial Internet.* 2015. URL: http://repositorio.cepal.org/bitstream/handle/11362/38767/S1500587_en.pdf.

Elvidge, C. D. et al. "Relation between satellite observed visible-near infrared emissions, population, economic activity and electric power consumption." In: *International Journal of Remote Sensing* 18.6 (1997), pp. 1373–1379. DOI: 10.1080/014311697218485. URL: http://www.tandfonline.com/doi/abs/10.1080/014311697218485.

Elvidge, Christopher D., Kimberly E. Baugh, et al. "Satellite inventory of human settlements using nocturnal radiation emissions: a contribution for the global toolchest." In: *Global Change Biology* (1997), pp. 387–395. URL: http://www.as.wvu.edu/biology/bio463/Elvidge%20et%20al%201997%20satellite%20night%20pictures.pdf.

Elvidge, Christopher D., Paul C. Sutton, et al. "A global poverty map derived from satellite data." In: *Computers & Geosciences* 35.8 (2009), pp. 1652–1660. URL: http://www.sciencedirect.com/science/article/pii/S0098300409001253.

Eurostat. *Big Data in Official Statistics: Technical Workshop Report*. 2014. URL: http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=102664009.

Expert Group on Poverty Statistics. "Expert Group on Poverty Statistics: Rio." In: *Compendium of Best Practices in Poverty Measurement*. 2006. ISBN: 85-240-3908-6.

Ginsberg, Jeremy et al. "Detecting influenza epidemics using search engine query data." In: *Nature* 457 (2009). DOI: `10.1038/nature07634`.

Giovannini, Enrico. "Statistics 2.0 - The next level." In: *10th National conference of statistics*. 2010. URL: `http://en.istat.it/istat/eventi/2010/10_conferenza_statistica/`.

Glickhouse, Rachel. *Explainer: Twitter in Latin America*. 2013. URL: `http://www.as-coa.org/articles/explainer-twitter-latin-america`.

GSMA Intelligence. *The Mobile Economy 2014*. URL: `http://www.gsmamobileeconomylatinamerica.com/GSMA_Mobile_Economy_LatinAmerica_2014.pdf`.

Guadamuz, Andés. "Habeas Data vs the European Data Protection Directive." In: *Electronic Law Journals* (2016). URL: `http://www2.warwick.ac.uk/fac/soc/law/elj/jilt/2001_3/guadamuz/`.

Gurin, Joel. "Big data and open data: what's what and why does it matter?" In: *The Guardian* (2014). URL: `http://www.theguardian.com/public-leaders-network/2014/apr/15/big-data-open-data-transform-government`.

Gutierrez, Thoralf, Gautier Krings, and Vincent D Blondel. "Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets." In: (2013). URL: `http://arxiv.org/pdf/1309.4496.pdf`.

Henderson, J. Vernon, Adam Storeygard, and David N Weil. "Measuring Economic Growth from Outer Space." In: *American Economic Review* 102.2 (2012), pp. 994–1028. ISSN: 0002-8282. DOI: `10.1257/aer.102.2.994`. URL: `http://pubs.aeaweb.org/doi/abs/10.1257/aer.102.2.994`.

Hubbard, Douglas W. *Pulse: the new science of harnessing Internet buzz to track threats and opportunities*. Hoboken, N.J: Wiley, 2011. 191 pp. ISBN: 978-0-470-93236-0.

Hyunyoung, Choi and Hal Varian. *Predicting the Present with Google Trends*. 2011. URL: `http://people.ischool.berkeley.edu/~hal/Papers/2011/ptp.pdf`.

Informa. *Latin America reaches 100% mobile penetration says Informa Telecoms & Media*. 2011. URL: `http://www.informa.com/media/press-releases-news/latest-news/latin-america-reaches-100-mobile-penetration-says-telecoms--media/`.

Instituto Nacional de Estadística y Geografía (INEGI) de México. *Agenda for International Meeting on Gender Statistics: Statistical challenges towards the implementation of the Post 2015 Agenda*. 2015. URL: `http://www.inegi.org.mx/eventos/2015/genero/doc/agenda_XVIgenero_en.pdf`.

International Telecommunication Union. *World Telecommunication/ICT Indicators database, 19th Edition*. 2015. URL: `http://www.itu.int/en/ITU-D/Statistics/Pages/publications/wtid.aspx`.

Internet World Stats. *Latin American Internet and Users and Population Statistics*. 2013. URL: `http://www.internetworldstats.com/stats10.htm`.

Katz, Raúl. *El ecosistema y la economía digital en América Latina*. 2015. URL: http://cet.la/blog/course/libro-el-ecosistema-y-la-economia-digital-en-america-latina/.

Khan, Amina and Elizabeth Stuart. *What's measured is also political*. 2015. URL: http://deliver2030.org/?p=5999.

King, Gary. "Big Data is Not About the Data!" In: *Golden Seeds Innovation Summit, New York City*. 2013. URL: http://gking.harvard.edu/files/gking/files/evbase-gs.pdf.

Kort, Eric A. et al. "Four corners: The largest US methane anomaly viewed from space." In: *Geophysical Research Letters* 41.19 (2014), pp. 6898–6903. ISSN: 00948276. DOI: 10.1002/2014GL061503. URL: http://doi.wiley.com/10.1002/2014GL061503.

Kreuter, Frauke and Roger D. Peng. "Privacy, Big Data, and the Public Good: Frameworks for Engagement." In: ed. by Julia Lane et al. Cambridge University Press, 2014. Chap. Extracting Information from Big Data: Issues of Measurement, Inference and Linkage, pp. 257–275. DOI: http://dx.doi.org/10.1017/CBO9781107590205.016. URL: http://ebooks.cambridge.org/chapter.jsf?bid=CBO9781107590205&cid=CBO9781107590205A020.

Letouzé, Emmanuel. "Six Considerations on Official Statistics and the (Big) Data Revolution." In: *Note prepared for the OECD–Paris21 event at the 2013 UN General Assembly, New York*. 2013.

— *Concept Note on SDGs and Big Data*. 2015.

Martinez-Herrera, Manuel. "From Habeas Data Action to Omnibus Data Protection: The Latin American Privacy (R)Evolution." In: *Latin American Law & Business Report* 19.9 (2011). URL: http://www.whitecase.com/sites/whitecase/files/files/download/publications/article_from_habeas_data_action_to_omnibus_data_protection.pdf2_.pdf.

Mejía, Luis Fernando et al. *Indicadores ISAAC: Siguiendo la actividad sectorial a partir de Google Trends*. 2013. URL: http://www.minhacienda.gov.co/portal/page/portal/HomeMinhacienda/politicafiscal/reportesmacroeconomicos/NotasFiscales/.

Mocanu, Delia et al. "The Twitter of Babel: Mapping World Languages through Microblogging Platforms." In: *PLOS ONE* 8.4 (2013), e61981. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0061981. URL: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0061981.

Montjoye, Yves-Alexandre de et al. "Unique in the crowd: The privacy bounds of human mobility." In: *Nature Scientific Reports* 3.1376 (2013). DOI: doi:10.1038/srep01376.

Newman Pont, Vivian. *Datos personales en informaci ón pública: oscuridad en lo privado y luz en lo público*. Dejusticia, 2015. ISBN: 978-9585885813.

Ohm, Paul. "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization." In: *UCLA Law Review* 57.1701 (2010). URL: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450006.

Open Data Institute. *The Data Spectrum helps you understand the language of data*. URL: https://theodi.org/data-spectrum (visited on 12/2015).

Open Data Research Network. *Opening Data in Montevideo: A bottom up experience*. 2014. URL: http://www.opendataresearch.org/content/2014/574/opening-data-montevideo-bottom-experience.

Orange. *Data for Development (D4D) Challenge*. 2014. URL: http://www.d4d.orange.com/.

Pawelke, Andreas and Anoush Rima Tatevossian. *Data Philanthropy: Where Are We Now?* 2013. URL: http://www.unglobalpulse.org/data-philanthropy-where-are-we-now.

Pentland, Alex "Sandy". "Social Computing and Behavioral Modeling." In: Boston, MA: Springer US, 2009. Chap. Reality Mining of Mobile Communications: Toward A New Deal On Data. ISBN: 978-1-4419-0056-2. DOI: 10.1007/978-1-4419-0056-2_1. URL: http://dx.doi.org/10.1007/978-1-4419-0056-2_1.

— "Reinventing Society in the Wake of Big Data: A Conversation with Alex (Sandy) Pentland." In: *Edge.org* (Aug. 30, 2012). URL: https://www.edge.org/conversation/alex_sandy_pentland-reinventing-society-in-the-wake-of-big-data (visited on 03/21/2016).

Pretz, Kathy. "Guadalajara: Smart City of the Near Future." In: *The Institute: The IEEE news source* (2014). URL: http://theinstitute.ieee.org/technology-focus/technology-topic/guadalajara-smart-city-of-the-near-future.

Reader, Ruth. *More than half of all smartphone users in Latin America use Twitter, study claims*. 2015. URL: http://venturebeat.com/2015/02/16/more-than-half-of-all-smartphone-users-in-latin-america-use-twitter-study-claims/.

Ruiz, Claudio. "Privacy and security, the Latin American way." In: *Digital Rights* 28 (2014). URL: http://www.digitalrightslac.net/en/privacidad-y-vigilancia-a-la-latinoamericana/.

Scannapieco, Monica et al. *Placing Big Data in Official Statistics: A Big Challenge?* Brussels, 2013. URL: http://www.cros-portal.eu/sites/default/files//NTTS2013fullPaper_214.pdf.

Schneising, Oliver et al. "Remote sensing of fugitive methane emissions from oil and gas production in North American tight geologic formations." In: *Earth's Future* 2.10 (2014), pp. 548–558. ISSN: 23284277. DOI: 10.1002/2014EF000265. URL: http://doi.wiley.com/10.1002/2014EF000265.

Secretaría de Turismo. *Uso Productivo de Big Data y Redes Sociales en el Sector Turismo*. 2014. URL: http://www.datatur.beta.sectur.gob.mx/Documentos%20Publicaciones/2014_1_DocInvs.pdf.

Smith, Christopher, Afra Mashhadi, and Licia Capra. *Ubiquitous Sensing for Mapping Poverty in Developing Countries*. 2013. URL: http://www.cities.io/wp-content/uploads/2012/12/d4d-chris-submitted.pdf.

Smith-Clarke, Christopher and Mashhadi, Afra and Capra, Licia. "Poverty on the Cheap: Estimating Poverty Maps Using Aggregated Mobile Communication Networks." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA, 2014. DOI: 10.1145/2556288.2557358. URL: http://doi.acm.org/10.1145/2556288.2557358.

Soto, Victor et al. "Prediction of Socioeconomic Levels Using Cell Phone Records." In: Springer Berlin Heidelberg, 2011, pp. 377–388. DOI: 10.1007/978-3-642-22362-4_35. URL: http://link.springer.com/10.1007/978-3-642-22362-4%7B%5C_%7D35.

Statistics Without Borders and Humanity Road. *Analysis of Twitter Data during Hurricane Sandy*. 2013. URL: http://www.slideshare.net/CatGraham/swb-hr-hurricane-sandy-twitter-analysis.

Sutton, Paul C., Christopher D. Elvidge, and Tilottama Ghosh. *Estimation of Gross Domestic Product at Sub-National Scales Using Nighttime Satellite Imagery*. 2007.

Téllez, Omar. "Producing Unicorns in The Land Of Fútbol, Samba and El Dorado." In: *TechCrunch* (2015). URL: http://techcrunch.com/2015/06/06/producing-unicorns-in-the-land-of-futbol-samba-and-el-dorado/.

The World Bank, World Bank Group, and Social Muse. *Big Data in Action for Development*. 2014. URL: http://data.worldbank.org/news/big-data-in-action-for-development.

Thenkabail, P S, N Gamage, and V U Smakhtin. "The Use of Remote Sensing Data for Drought Assessment and Monitoring in Southwest Asia." In: *International Water Management Institute* (2004). URL: http://www.iwmi.cgiar.org/Publications/IWMI%7B%5C_%7DResearch%7B%5C_%7DReports/PDF/pub085/RR85.pdf.

Tufekci, Zeynep. *The year we get creeped out by algorithms*. 2014. URL: http://www.niemanlab.org/2014/12/the-year-we-get-creeped-out-by-algorithms/.

United Nations Economic Commission for Europe (UNECE). *What does "Big Data" Mean for Official Statistics*. 2013. URL: http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170614.

United Nations Global Pulse. *Daily Tracking of Commodity Prices: The E-bread Index*. 2011. URL: http://www.unglobalpulse.org/projects/comparing-global-prices-local-products-real-time-e-pricing-bread.

— *Rapid Impact and Vulnerability Analysis Fund (RIVAF) Final Report*. 2012. URL: http://www.unglobalpulse.org/sites/default/files/FINAL%20RIVAF%20REPORT%20COMPILED_0.pdf.

— "Mining Indonesian Tweets to Understand Food Price Crises." In: (2014). URL: http://www.unglobalpulse.org/sites/default/files/Global-Pulse-Mining-Indonesian-Tweets-Food-Price-Crises%20copy.pdf.

United Nations Statistical Commission. *Big data and modernization of statistical systems*. 2014. URL: http://unstats.un.org/unsd/statcom/doc14/2014-11-BigData-E.pdf.

United Nations Statistical Commission. *Report of the Global Working Group on Big data for official statistics*. 2014. URL: `http://unstats.un.org/unsd/statcom/doc15/2015-4-BigData.pdf`.

Urioste Braga, Fernando. *Derecho de la información*. Montevideo-Buenos Aires: B de F, 2009.

Velasco, Patricio and Pablo Viollier. “Información Financiera y Discriminación Laboral en Chile: un Caso de Estudio Sobre.” In: *Derechos Digitales* (2016). URL: `https://www.derechosdigitales.org/wp-content/uploads/big-data-informe.pdf`.

Yuan, Q. and Et al. “Monitoring Influenza Epidemics in China with Search Query from Baidu.” In: *PLOS ONE* 8(5): e64323 (2013).

Zhang, Renhua et al. “Drought Monitoring in Northern China based on Remote Sensing Data and Land Surface Modeling.” In: *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. Vol. 3. 1. IEEE, 2008, pp. III – 860–III –863. ISBN: 978-1-4244-2807-6. DOI: `10.1109/IGARSS.2008.4779485`. URL: `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4779485`.